DOCUMENT RESUME

ED 157 996 CE 016 446

AUTHOR Wiley, Llewellyn N.; Hahn, Clifford P.

TITLE Task Level Job Performance Criteria Development. Final Report.

INSTITUTION Air Force Human Resources Lab., Brooks AFB, Tex. Occupational and Manpower Research Div.; American Institutes for Research in the Behavioral Sciences, Washington, D.C.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-77-75
PUB DATE Dec 77
GRANT F41609-71-C-0010
NOTE 59p.; Parts of document may be marginally legible due to small type

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Aviation Mechanics; Enlisted Personnel; Equipment Maintenance; *Evaluation Methods; *Job Analysis; *Job Skills; Peer Evaluation; *Performance Criteria; Predictor Variables; Radio Technology; Self Evaluation; Statistical Analysis; Tables (Data); *Task Performance; Telecommunication

ABSTRACT

This study investigated possibilities for improving identification of job performance requirements by examining job incumbent's performance on separate tasks. Three specialties were studied: 291X0, Telecommunications Operation Specialist; 304X4, Group Radio Communications Equipment Repairman; and 431X1C, Aircraft Maintenance Specialist. Incumbent's task performance was rated by incumbents, peers, and supervisors; and tests and inventories (eleven cognitive tests, Biographical Inventory, Vocational Interest Career Examination, and Forty-three Item Job Satisfaction Information blank) were administered. Air Force records were obtained to secure the following data: incumbent grade, service time, sex, education, and Aptitude Index scores. Two correlations were run between raters: correlating performance on separate tasks and correlating performance on six appraisal dimensions. Cross-rater reliabilities were low, but significant, on task assessments, and in the r=.40 range on appraisal. Low correlations were found for nontask predictors such as grade, service time, and Aptitude Indexes. All measures, except training origins and task performance satisfaction, were put into regression problems to account for peer and supervisor ratings. Data suggest different factors were important for different kinds of work and for different performance appraisal dimensions. Difficult tasks (in terms of learning time) were better measured on performance. (Author/CSS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-77-75

AIR FORCE HUMAN RESOURCES LABORATORY

ED157996

TASK LEVEL JOB PERFORMANCE CRITERIA DEVELOPMENT

By

Llewellyn N. Wiley

OCCUPATION AND MANPOWER RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

Clifford P. Hahn
American Institutes for Research
1055 Thomas Jefferson St., N.W.
Washington, D.C. 20007

December 1977

Approved for public release; distribution unlimited.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CE 016 446

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by American Institutes for Research, 1055 Thomas Jefferson St., N.W., Washington, D.C. 20007 under contract F41609-71-C-0010, project 7734, with Occupation and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

RAYMOND E. CHRISTAL, Technical Director
Occupation and Manpower Research Division

DAN D. FULGHAM, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-77-75
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): TASK LEVEL JOB PERFORMANCE CRITERIA DEVELOPMENT
5. TYPE OF REPORT & PERIOD COVERED: Final
6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Llewellyn N. Wiley, Clifford P. Hahn
8. CONTRACT OR GRANT NUMBER(s): F41609-71-C-0010

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Occupation and Manpower Research Division, AFHRL, Brooks AFB, Texas 78235, and American Institutes for Research, 1055 Thomas Jefferson St., N.W., Washington, D.C. 20007
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 77340601, 77340602

11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235
12. REPORT DATE: December 1977
13. NUMBER OF PAGES: 54

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The contractor's final report (Hahn, 1975) covered the collection of the data and analyses up to a point, which is defined in the report. Terminal analyses were performed by the Occupation and Manpower Research Division, of AFHRL, and the data constitute a reservoir for analyses under 77340602.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

AFSC 291X0	demographic predictors	peers' overall ratings
AFSC 304X0	incumbents' task ratings	peers' take ratings
AFSC 431X1C	overall performance dimensions	performance ratings reliability
aptitude predicting performance	overall performance rating	self-performance ratings
comparing AFSC performance	performance patterns	supervisors' overall ratings

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study investigated the possibilities for improving the identification of the requirements for jobs by studying performance of job incumbents on separate tasks. Three specialties were selected for study: 291X0, Telecommunications Operations Specialist; 304X4, Ground Radio Communications Equipment Repairman; 431X1C, Aircraft Maintenance Specialist, single- and dual-engine jet. Incumbents, peers, and supervisors rated the performance of the incumbents on a selected set of tasks. In addition, job inventories and an experimental test battery were administered to the incumbents. The battery included 11 short experimental cognitive tests, a Biographical Inventory, the Vocational Interest-Career Examination (VOICE), and a 43-item Job Satisfaction Information blank. Data of record were also obtained from Air Force files to provide such items as incumbent grade,

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 19 Continued:

supervisors' task ratings
task difficulty effects
task performance dimensions
task performance ratings
task predicting overall
task rating reliability

Item 20 Continued:

service time, sex, education at enlistment, and Aptitude Index scores. Correlations were run between raters, correlating performance on separate tasks, and between raters, correlating performance on 6 overall dimensions of appraisal. Cross-rater reliabilities were low, but significant, on task assessments, and in the r = .40 range on overall ratings. Similarly low correlations were found for nontask predictors, such as grade, service time, and Aptitude Indexes. All types of obtained measures, except data on the origins of training and on task performance satisfaction, were put into regression problems to account for the 6 overall performance ratings made by peers and supervisors. The data suggest that different factors were important for different kinds of work, and for different dimensions of performance appraisal. Of all the many findings of the study, by far the most enlightening was that difficult tasks (in terms of learning time) were better measured on performance. This arose from less use of the top of the rating scale, and it produced lower performance appraisals from the group (AFSC 304X4) which had been selected by the Air Force for having the highest aptitude scores. Should subsequent analyses prove that this finding also applies to job ratings within AFSCs, the result would have implications for Air Force job performance appraisal.

PREFACE

This study was made possible by the original funding arranged through the Office of the Secretary of Defense for Manpower and Reserve Affairs. The proposal and statement of work were written by Dr. Raymond E. Christal, assisted by Mr. William B. Lecznar and Dr. Llewellyn N. Wiley, who successively became the contract monitors, 1971–1973, and 1973–1974.

The authors are grateful to Dr. Robert W. Stephenson for his considerable input to the organization of this report, and for his contribution to the analyses.

Special thanks are due to Mr. William N. Wallace for his work in the administration and follow-up of all survey materials, the supervision of the monumental data transcription task, and the conduct of preliminary data analyses. Thanks are also due to Dr. Paul W. Fingerman and Dr. Robert Frey for the conduct of computer analyses; Dr. Charles A. Darby and Dr. David I. Sheppard for their work in the development of preliminary forms and scales; and Ms. Joan E. Wallace for her work in the development and early field testing of survey materials and procedures.

Data analyses accomplished by AFHRL were programmed by Mr. James L. Friemann, A1C Stanley E. Prescott, and Sgt T. G. Smith; and the data were processed by Mr. Lewis A. Walker, Mr. James L. Brazel, and A1C Ronald J. McCary.

We are also endebted to Mr. Kenneth D. Koym, Mrs. M. Joyce Giorgia, and Mrs. Mary D. Spencer for AFHRL's part in the collection and preparation of the data. Mr. Kenneth Finstuen of AFHRL, also assisted in the final processing of this report.

TABLE OF CONTENTS

	Page
I. Introduction	5
II. Background	6
III. Development of Performance Measurement Instruments and Procedures	8
Data Sources	9
IV. The Survey Materials	9
Peer and Supervisor Performance Rating Booklet	10
The Experimental Test Battery	10
Some Specific Comments on the Survey Materials	10
V. Sample Sizes	10
VI. Description of Three Samples	12
VII. Analyses of Unrestricted Samples	12
Distributions of Task Performance Ratings	12
Skill and Ability Versus Motivation Ratings	16
Correlations of Task Performance Ratings Made on the Same Incumbents	17
Prediction of Task Performance Ratings from Data of Record and Cognitive Test Scores	20
Correlations Among the Six Overall Performance Rating Dimensions	22
Correlations Between Task Performance Ratings and Overall Performance Dimension Ratings	24
Summary of Findings for Unrestricted Samples	24
VIII. Regression Analyses of Restricted Samples	27
The "Flagged" Sample Concept	27
Combining Raters	28
Prediction from Data of Record	29
Reduction of Five Task Performance Predictors	34
Cross-Rater Prediction of Overall Performance	41
Contributions of Single Variables of Data of Record	41
IX. Summary of Regression Problems	44
X. Discussion and Conclusions	45
XI. Summary	48
References	50
Notes and Study Numbers	51

LIST OF ILLUSTRATIONS

Figure		Page
1	Sequence of the contract effort	6
2	Comparative use of scale values in task ratings	13
3	Distributions of mean task ratings	14
4	Peer vs. supervisor task performance rating correlations	19
5	Correlations of task and general performance ratings, product	25
6	Correlations of task and general performance ratings, motivation	26
7	Comparative Task Difficulty Means by AFSC	46

LIST OF TABLES

Table		Page
1	Selected Samples, the "Flagged Population"	11
2	Significant Correlations with Task Performance Ratings	21
3	Correlations Among Six Overall Performance Rating Dimensions	23
4	Task Performance Ratings Predicting Overall Performance	28
5	Correlations with Performance Ratings AFSC 291X0	29
6	Correlations with Performance Ratings AFSC 304X4	30
7	Correlations with Performance Ratings AFSC 431X1C	30
8	Regression Problems Compared for Three AFSCs – R^2s	31
9	Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 291X0	32
10	Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 304X4	33
11	Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 431X1C	33
12	Statistics of Selected Tasks	35
13	Unqiue Prediction Contribution of 5 Task Ratings to Overall Performance	36
14	Unique Demographic Contribution to Aptitude and 5 Task Overall Performance Prediction	37
15	Unique Aptitude Contribution to Demographic and 5 Task Overall Performance Prediction	38
16	Demographic Contribution to Performance Prediction by 5 Tasks	39
17	Aptitude Contribution to Performance Prediction by 5 Tasks	40
18	Cross-Rater Regression Contributions from Sets of 5 Task Performance Ratings	42
19	Comparison of Single Variable Contributions to Overall Prediction by Task Ratings	43

TASK LEVEL JOB PERFORMANCE CRITERIA DEVELOPMENT

I. INTRODUCTION

This is an account of the major steps and findings in the execution of Contract F41609-71-C-0010, secured by the Occupational Research Division of the Air Force Human Resources Laboratory (AFHRL), Brooks Air Force Base, Texas, and performed by the American Institutes for Research (AIR), Washington, DC, during the period 1971 through 1974. The basic document for a large portion of this account is the final management report, having the same title, authored by the Principal Investigator, Mr. Clifford P. Hahn. Extensive extracts have been made from that report and condensed for presentation here. The contract report and its appendices are available from the Defense Documentation Center (Hahn, 1975).

Perhaps no better explanation of the aims of the study can be had than to quote the Introduction and Background section of the contract proposal of 16 April 1970.

> One of the greatest needs of managers of the military manpower and personnel systems is for a method to accurately measure how well individuals perform on the job. The Personnel Research Division (AFHRL) has developed techniques by which the Air Force can determine the tasks and jobs being assigned to personnel; but little or no information is available concerning how well these tasks and jobs are being performed. Official supervisory ratings do not serve this purpose well enough. Such ratings are global in nature, not specifically related to tasks and jobs, highly inflated, and provide insufficient variance for discriminating among individuals being rated.
>
> The upshot is that selection and classification devices are designed to maximize performance in school, rather than performance on the job. Training courses are presumably tailored to job content, but adequate procedures are not available to determine their efficiency. Proficiency tests contain questions about tasks likely to be encountered, but are not validated against job performance. In short, there is no way to demonstrate that individuals with high aptitude scores, who have undergone extensive formal training, and who score high on proficiency tests, actually perform significantly better on the job than individuals having lower aptitude scores and less training, and scoring lower on proficiency tests. Until better criteria are available, it will be difficult to evaluate new selection devices, training techniques, occupational structures, assignment procedures, classification models, or a host of other management programs and devices.
>
> There appears to be little hope of obtaining good performance ratings from supervisors as long as such ratings are (a) revealed to the ratee and (b) used for making decisions about the ratee's promotions. Nor is it feasible to develop objective performance tests by which subordinates can be evaluated by unbiased observers in a controlled setting. There are approximately 30,000 tasks performed by Air Force enlisted personnel. It would not be feasible to construct and administer performance tests except in a few critical areas. It is also recognized that how an individual behaves for a short period of time, when he knows he is being closely observed, does not necessarily correlate with how well he behaves in the operational environment, when he is not being evaluated closely. Thus, one is thrown back to obtaining performance information from those who are in a position to observe workers in an operational setting.

The foregoing is sufficient to sketch the magnitude of the problem and to identify the many applications that the data might have if they were available. It is not a problem that has been peculiar to the Air Force, and it was viewed with such concern that initial funding of the contract was provided by the Department of Defense.

The task rating approach to job performance was designed to provide information that cannot be had when performance is assessed on a global basis. It was known that jobs differ greatly with respect to the tasks comprising them, although the Air Force specialty codes (AFSC) held by the incumbents in those jobs might be identical. If it were possible to reduce performance measurements to elements (tasks) of jobs reliably and with agreement among judges, the relative importance of many factors contributing to overall performance might be revealed. In particular, it might be possible to weigh the contribution of an incumbent's aptitude, experience, and attitude toward establishing an overall judgment regarding his performance. If both the task ratings and the overall ratings were made by the same judges, control would be exercised over sources of error. It would not be necessary to make suppositions regarding the

equivalence of task performance raters in one sample and overall performance raters in another. If two or more specialties were studied, it would be possible to compare them with respect to the way performance factors assembled themselves to yield overall assessments.

II. BACKGROUND

A number of ground rules were set forth in the work statement of the contract, most of which could be followed to the letter, but some of which proved to be infeasible in the operational situation. An outline of the full sequence of contract and analysis events is given as Figure 1.

Agency	Activity
AFHRL	Prepares statement of work and awards competitive bid to AIR.
AIR	Procures information on AFSCs from files and narrows specialty list.
AIR	Begins base interviews with NCOs to select specialties.
AIR	Continues base visits and prepares materials for field reviews. Narrows list to three specialties, selects coordinators, and prepares materials for assembled meetings with 50 NCOs per specialty.
AIR	Hosts a week-long meeting with coordinators for each AFSC at Lackland AFB, Texas. Tries out preliminary scales after reviewing selected task items and performance dimensions. Sends materials to printer.
AFHRL	Selects the experimental test battery, with AIR inputs to items. Provides base rosters for surveys.
AIR	Mails out survey materials to base coordinators via Consolidated Base Personnel Offices. Produces testing manual.
AFHRL	Procures Survey Control Number and prints test battery. Sends test batteries through Test Control Officers. Scores returned tests and prepares card image tape of test data for AIR.
AIR	Records survey responses and makes initial report on number of cases collected.
AFHRL	Decides to perform supplementary survey and selects materials and survey procedures, using normal channels. Provides additional case rosters, and has returned surveys directed to AIR.
AFHRL	Obtains address list of personnel surveyed at Time 1 and still in the Air Force at Time 2. Obtains Time 2 ratings for direct delivery to AIR. Scores supplemental survey tests and sends card image tape to AIR.
AIR	Delivers original survey data tape file to AFHRL.
AFHRL	Runs shakedown tests on AIR data tape. Performs studies of interrater reliabilities paralleling AIR analyses. Classifies verbal comments made in peer and supervisor reports.
AIR	Performs correlations among task and overall ratings; studies aptitude/skills vs. motivation ratings.
AFHRL & AIR	Hold coordination meeting on final analyses, deciding to eliminate additional analyses of incumbents as performance raters, and to limit computed means to two raters. Establish a limited set of data of record predictors, covering all essential demographic variables.
AIR	Delivers zero-order correlations of variables and first portion of final report. Contract ended 31 December 1974.
AFHRL	Delivers to AIR tape file of predictor variables after matching cases with Air Force data record files. Runs regression analyses to see if suitably scored attitude and satisfaction tests could predict both task and overall performance ratings. Determines overall rating reliability with cross-rater regressions. Charts future analyses by using graphical interpretations.
AIR	Delivers data and correlation printouts with additional final analyses, ending final report. Delivers Time 2 data tapes.
AFHRL	Continues regression analyses, setting up the "flagged sample" procedure. Analyzes relative contributions of task ratings and data of record. Selects techniques to reduce the number of predictors, and tests contributions of aptitude and demographic data against just 5 task performance ratings as predictors. Runs analyses to determine if rater tendency was cause of findings. Sets up files and begins analyses to answer questions not covered by the report. Writes the report.

Figure 1. Sequence of the contract effort.

Considerable effort was expended upon selection of the specialties to be used in the study. The aid of noncommissioned officer (NCO) consultants was obtained from Major Commands, and visits lasting up to a week were made to bases. This ended with the choice of three AFSCs: 29150, Telecommunications Operations Specialist; 30454, Ground Radio Communications Equipment Repairman; and 43151C, Aircraft Maintenance Specialist (single- and dual-engine jet). The choice of specialties fulfilled a number of contract stipulations aimed at collecting data that would represent the Air Force broadly. Among the conditions required were adequate numbers of job incumbents, presence of both hardware and software specialty activities, availability of bases for direct visitation, availability of a current job inventory in the specialty, and at least one specialty that included personnel who did not receive formal technical school training. It chanced that all three chosen specialties operated in shifts. Only the communications center sample contained women, practically all of whom were telephone switchboard operators.

Selection of the specialties was accomplied by, and followed by, development of draft form rating instruments and the choice of an initial set of tasks for tryout. Approximately 50 NCOs in each AFSC were convened by the contractor for a week's experimental conference and workshop that were held successively at the Lackland AFB, Texas, facility of AFHRL. At this time, the NCOs were indoctrinated in performance evaluation and observation. Then they used the instruments and discussed the observability of the selected tasks. Problems and facts pertinent to task performance ratings were brought out in each specialty, resulting in separate rating dimensions for certain tasks. The usefulness of the ratings for future Air Force decisions was emphasized, and the NCOs were returned to their bases to act as focal points for indoctrinating additional raters, and to coordinate the surveys to follow. It was here that difficulties arose. The survey control number (SCN) system was introduced into the Air Force while the survey forms were being printed, and these emerged unnumbered. No surveys without a control number were authorized, and the 1972 surveys could not be authorized by base personnel offices without a control number. During the resulting delays some of the chosen NCOs were transferred and participant interest in the study waned. Data returns were slow, unpredictable, often incomplete, and smaller than anticipated.

The job incumbent was asked to complete a standard Air Force job inventory, a booklet in which he rated himself on the performance of those tasks he did in the selected set, a booklet in which he indicated his source of knowledge in or experience at each task he performed, and another booklet indicating the type of satisfaction or annoyance he had from performing each task. The incumbent also rated 10 job factors covering his whole job. No exact check was made to determine how many of these incumbents also acted as peer raters.

Since any specific task was performed by less than half the individuals in most samples, and some tasks were performed by very few incumbents, it was essential to have fairly large groups in order to have stable data at the task level. There was, for example, no way of assuring that two raters would rate an incumbent on the same tasks, although the probability of their doing so was high. As the difficulties of data collection developed, it was feared that the major objectives of the contract would not be met without additional sampling. In the fall of 1973, AFHRL supplemented the AIR data by sending out additional surveys and a battery of selected tests. These were handled through the test control officers, and the data recovery improved. In the supplemental survey, incumbents did not fill out job inventories nor the work factor ratings; but they were asked to complete the three survey forms and to take the test battery. Besides the incumbent, only one rater, a supervisor, was used.

Time 2 surveying of previously rated incumbents began in early 1974. All incumbents available for a year were located, and their bases were requested to have them rated by two supervisors, wherever possible. Incumbents did not participate and were not informed.

In the spring of 1974, the complete data tapes of the Time 1 surveys were forwarded by the contractor to AFHRL and analyses were begun at both places. AFHRL performed both parallel and independent analyses to investigate features of the data of future interest, as well as those of immediate concern. In October 1974, an analysis planning conference was held to determine which analyses were essential for the contractor to perform with existing funds. (It had never been planned to exhaust the data

reservoir to provide the contract final report, since it was evident that many tangential problems would prove to be interesting.) The present report extends data analysis activities beyond those of the contractor in order to achieve closure on findings regarded as having major interest for the Air Force. Some additional work that was performed subsequently by AFHRL is reported elsewhere (Wiley, 1976).

III. DEVELOPMENT OF PERFORMANCE MEASUREMENT INSTRUMENTS AND PROCEDURES

This section is quoted in full from the contractor's management report, pages 6 through 9 (Hahn, 1975).

> A major task in developing the criterion instruments and procedures sought for field use involved translating the descriptive task statements of the USAF Job Inventories into evaluative statements that could form the basis for scaling how well the tasks were done. Throughout the entire development process, contractor staff members who were experienced in the techniques of performance evaluation and scaling had the active support of a group of experienced incumbents from the career ladders being studied. This was accomplished through an interative series of working sessions, first with groups of 2 to 3 incumbents per ladder, then groups of 10, and finally groups of approximately 50 incumbents for each career ladder.
>
> The Job Inventory task lists were reviewed and appropriate additions, selections, and revisions were made in order to update the task lists. Data from iterative sorts of discernible levels of performance; degree of observability and measurability; criticality; range of difficulty; and stability of performance were used to develop an initial list of candidate tasks for further development. Career ladder incumbents then developed an initial set of behavioral descriptors for total task performance or for critical dimensions of performance for the tasks on the candidate list.
>
> A series of one-week work sessions with groups of 10 experienced incumbents were conducted to review and revise the descriptors previously developed, to develop additional descriptors for relevant tasks, and to devise initial scaling procedures for use with the descriptors. These same groups rated the importance of each task dimension for inclusion in the field survey forms. The data from these sessions were used by the contractor staff in preparing forms for use in simulated rating sessions by larger groups of NCOs from each career ladder.
>
> A series of week-long ladder workshops was held, each involving approximately 50 senior NCOs from the career fields being studied. The purposes of these workshops were to use previously developed instruments in a simulated rating situation, to develop additional scales, to develop ancillary instruments for field use in conjunction with the performance rating instruments, and to elicit opinions about procedures to be used to collect the field data.
>
> Results from the career ladder workshops indicated that the forms developed appeared capable of capturing some of the performance variance that existed in the field. Six overall job performance ratings were generated in addition to the ratings at the task level, as were judgments of importance for inclusion of the surviving tasks in the final field format. Data from the workshops were utilized to make final revisions in the field forms for collecting performance data and these were later reviewed by workshop participants on a mail-out basis.
>
> Several additional forms were developed and tried out in the career ladder workshops. These forms were to be utilized in connection with the performance data survey to satisfy other concerns of interest to the monitoring agency. One of these was a Work Factor Requirements Rating Form. The work factors concerned were those which applied to the job itself and not to the airmen performing the job nor the manner in which it was performed. These were the types of factors that are typically considered during job evaluation procedures designed to establish an appropriate grade and pay level. It was anticipated that such factors would eventually be compared with performance data. A ten-factor form with a nine-point scale for rating each factor was developed for field use.
>
> In order to understand better some of the factors which contribute to task performance, instruments and procedures were also developed to collect data regarding the acquisition and retention of the skills and knowledges associated with various tasks. The form developed called for judgments of the major source of skill acquisition in terms of technical training school, a formal OJT program, or job experience. Separate ratings were requested for acquisition of job knowledge and job proficiency following the model of the Air Force dual-channel OJT concept. Judgments were also requested concerning the relative perishability of task knowledge and proficiency after an acceptable level had once been attained.
>
> Tentative forms were also developed for judging the interest value and judged complexity of the various tasks. These ratings tended to be highly intercorrelated and their relationship to overall performance was somewhat unstable across the three career ladders. Data regarding these task characteristics were therefore not sought from the field.

Tentative forms were also developed for obtaining task preference ratings on the premise that individual preferences for certain tasks or groups of tasks might affect the motivation level and thus influence task performance. For tasks selected as most and least preferable, judgments of the relative potency of generalized motivational factors were requested. The data from these activities were used to prepare a motivation rating form to be completed by incumbents. This form allowed for expressions of both the importance of the motivational factors and the direction of their influence; i.e., positive, negative, or both.

Data Sources

As a result of the developmental activities described, the following set of survey instruments was utilized for collecting data for 5-skill-level incumbents from the three career ladders.

1. *Performance and Skills/Abilities Versus Motivation Ratings.* This rating instrument was designed for use by both supervisors and peers to provide data on the level of task dimension performance and on the relative importance of skills and abilities as opposed to motivational factors in contributing to the level of performance.

2. *Performance Ratings.* This rating instrument was designed for use by incumbents to provide data concerning their own perceived level of task dimension performance.

3. *Motivation Ratings.* This rating instrument was designed for use by incumbents to provide data regarding both the intensity and direction of effects of a set of generalized motivational factors.

4. *Training and Skill Retention Ratings.* This rating instrument was designed for use by incumbents as well as by supervisors and peers to provide data regarding the primary source for acquisition of task knowledge and task proficiency, as well as the relative perishability of such knowledge and proficiency after an acceptable level had once been attained.

5. *Work Requirement Factor Ratings.* This rating instrument was designed for use by incumbents, supervisors, and peers to provide data on ten generalized requirement factors associated with the duty positions within the career ladder themselves rather than with incumbents or their level of performance.

6. *United States Air Force Job Inventory.* Copies of the current job inventory were reproduced and used by incumbents for indicating which tasks they performed and the relative amount of time spent on each.

IV. THE SURVEY MATERIALS

General instructions aimed at motivating respondents were provided to all participants in the surveys, and each booklet contained a specific set of instructions. Correlative data were obtained by matching incumbent names and social security numbers against Air Force personnel record tapes maintained by AFHRL. While the matched data were not part of the survey, much of the matched data figured prominently in the analyses. In addition to the general orientation provided to all participants, peers and supervisors were given tips about observing and rating job performance.

Besides completing job inventories, incumbents rated themselves on task performance in the first survey booklet, reported their reactions to performing the same tasks in the second booklet, and in the third booklet gave information regarding their sources of training and their retention of skills. They indirectly appraised the requirements of their jobs by rating work requirement factors.

Peers and supervisors rated specific incumbents on task performance, by task, and using a booklet identical to the incumbent's, rated the tasks on sources of training and skill preservation.

Those incumbents in the initial survey who complied with all the requested activities completed the following:

1. A current Air Force job inventory in their ladder
2. A self-rating booklet of task performance dimensions (7-point scale)

3. A motivation-supplied-by-task rating booklet
4. A task training source and skill retention booklet
5. An experimental test battery, containing, 11 short coginitve tests, a biographical inventory, a 400-item Vocational Interest-Career Examination (VOICE), a least-preferred-coworker set of scaled items, and, a 43-item list of job satisfaction determiners to be rated
6. A single rating page of 10 work requirement factors.

Peer and Supervisor Performance Rating Booklet

The peer and supervisor performance rating booklet came next, with the overall rating page, which is part of the same booklet. It was preceded by information concerning tips about observing and rating job performance.

The peer and supervisor performance rating booklet contained information identifying the ratee, his job title, grade, skill level, the length of time he was known by the rater, estimates of the amount and kind of contact the rater had with the ratee, and provided space for general comments.

The Experimental Test Battery

The experimental test battery was administered by AFHRL. Task scores for these cognitive tests are included in Table 1. Also included in the battery were the Biographical Inventory and VOICE, the Vocational Interest Career Examination. These were followed by the Least Preferred Coworker rating instrument and a 43-item Job Satisfaction Information List. (The biographical material would have to be updated extensively for current use, and the VOICE instrument is in the process of refinement by AFHRL.)

Some Specific Comments on the Survey Materials

A single item of the 43-item job satisfaction list, the one concerned with how the Air Force meets its commitments to the individual, was of strictly temporary interest. This item reflected attitudes existing toward Air Force service in 1972, rather than attitudes toward the incumbent's specific job. Graphical analyses showed that the majority of the responses were unfavorable, and this was true for all three AFSCs. The responses were the most extremely negative of all 43 items in the list.

Other attitude and interest items have been grouped for coding and weighting purposes, then examined graphically. Interesting comparisons among the three AFSCs resulted. To cite but one of these, the graphs showed that aircraft mechanics frequently considered their working conditions to be poor, but that their general satisfaction with their job assignments was predominantly better than that shown by personnel in the other two AFSCs.

The instructions for the incumbent task motivation booklet may have presented the incumbent with a hard problem, if he was conscientious in his responses. Since it is impossible to know whether or not the incumbent understood these complicated instructions, or was responding in any but a perfunctory manner, it is doubtful that definitive analyses can be made of the motivation booklet responses.

Analyses will be presented with only part of the incumbent's possible contribution – his self-ratings on task performance, and his actual performance on tests. The analysis of peer and supervisory data will be primarily concerned with performance ratings, and they cannot utilize information about job relations with the ratee.

V. SAMPLE SIZES

Rather than list all the types of data combinations, certain classes of data have been analyzed for the largest possible N. When intercorrelations are being computed, the matrices must be complete, and the data

Table 1. Selected Samples, the "Flagged Population"

Identifying Variable	AFSC 291X0 N = 457 Mean	AFSC 291X0 N = 457 SD	AFSC 304X4 N = 399 Mean	AFSC 304X4 N = 399 SD	AFSC 431X1C N = 487 Mean	AFSC 431X1C N = 487 SD
Demographic Data of Record						
Grade (1–9)	4.118	.648	4.150	.706	4.528	.558
Months, tot act Fed mil service	55.821	46.084	52.331	37.630	58.983	21.973
Sex male = 1 / other = 0	398′	.335	399′	.000	487′	.000
Age in months at enlistment	292.068	47.764	290.088	38.884	293.655	26.156
Educational level at enlist.	4.112	.457	4.135	.550	4.019	.422
Size of city of origin (1–5)	2.626	1.316	2.609	1.287	2.448	1.280
Married = 1 / other = 0	257′	.496	233′	.493	355′	.443
Divorced or separated = 1 / 0	3′	.081	4′	.100	9′	.135
Single = 1 / other = 0	197′	.495	162′	.491	122′	.433
Aptitude Data of Record, Aptitude Indexes						
Mechanical AI	47.269	21.558	77.043	15.070	63.737	13.029
Administrative AI	71.543	13.443	75.627	15.535	51.306	19.257
General AI	63.260	17.011	78.922	13.691	55.934	17.016
Electronic AI	58.107	19.027	85.952	9.569	58.809	17.365
Experimental Cognitive Test Scores						
Decoding	19.050	12.448	25.501	11.058	18.875	11.840
Memory for landmarks	16.042	9.309	20.378	8.769	14.694	8.548
Complex scale reading	4.125	3.099	6.652	4.363	4.780	3.545
Pursuit	29.234	10.705	32.343	10.588	29.869	12.361
Figure analogies	20.449	7.968	26.566	5.605	19.975	8.450
Hands	28.630	17.554	36.153	13.075	33.187	14.610
Cubes	14.230	6.514	19.333	6.425	16.039	6.528
Mechanical principles	34.184	15.073	57.384	18.479	44.522	17.493
Following directions	25.179	18.964	34.947	18.528	23.528	18.442
Practical estimations	7.619	3.738	9.797	4.003	8.290	3.701
Spatial reasoning	5.212	8.648	12.466	10.605	3.943	8.416
Overall Performance Ratings By Peers and/or Supervisors Combined						
General performance	82.065	13.177	76.426	15.357	80.491	13.589
Amount of work performed	82.565	14.406	75.265	16.500	80.360	14.595
Quality of work	83.875	14.154	79.220	15.523	83.686	13.515
Will do more than his share	81.374	17.636	74.184	20.487	77.644	19.423
Self-initiating	83.435	16.291	75.044	20.196	79.767	18.735
Will share knowledge	84.082	15.013	80.015	16.484	82.923	16.181

′Converted to actual n.

Note. — The educational scale used in Table 1 is as follows:

1 = elementary school, graduated or not
2 = high school, one through three years
3 = completed high school equivalency tests but has no diploma
4 = high school graduate
5 = one or more years of college, includes an AA degree, or a graduate of a diploma school program, including registered nurse, but not a 4-year college diploma
6 = one or more bachelor's degrees, includes optometry and podiatry
7 = master's candidate under USAFIT
8 = master's degree and above in anything less than a doctorate
9 = all earned doctor's degrees, also LLB and JD

have been stripped down for all cells to the same types of cases within each AFSC. In the one instance of task performance ratings used in regression problems, special matrices had to be constructed to provide the same number of entries in all cells. The minimum N for any sample providing all types of data used in these analyses appears for those regression problems in which complete data were demanded for overall ratings, test battery scores, and personnel data. The three Ns in that instance were 457 for AFSC 291X0, 399 for 304X4, and 487 for 431X1C. (These samples appear in Table 1.) Subsample Ns for task performance ratings range from the very small to the hundreds.

VI. DESCRIPTION OF THREE SAMPLES

To describe the population of this study, one is faced with the alternative of supplying the means and sigmas for the largest sample measured on each variable or of selecting a group of incumbents whose membership contains only individuals with data on every variable. The latter choice was made after comparing means and standard deviations of larger populations to determine if the sample statistics were representative of the population at large. Table 1 gives the means and standard deviations of the three incumbent samples used in regression problems to determine the relative contribution of various kinds of measures toward accounting for the six overall performance ratings.

The table means and standard deviations supply striking facts about incumbents in AFSC 304X4: (a) they were slightly younger chronologically and in service time than members of the other two groups. (b) they entered the Air Force slightly better educated than the others, (c) they excelled the other two groups on all Aptitude Indexes, and (d) they excelled the other two groups on every cognitive test; but, their six performance ratings were, dimension-by-dimension, lower than those of the other two groups.

The only significant characteristic of the 291X0 AFSC sample shown in the table is that 59 of the 457 incumbents were female, the only women in the study. In general, the 291X0 AFSC performance ratings were the highest. The performance ratings of the AFSC 431X1C sample showed the greatest variability among the performance rating dimensions. Compared to the data revealing the high aptitude and lower performance rating of incumbents in AFSC 304X4 these items are trivial.

Variables relating to time on base and job, which were employed by the contractor, were not available for participants in the supplemental survey, and were thus not available for the three selected samples in all cases. Task rating data are represented sporadically throughout the study and cannot be provided in a summary table.

VII. ANALYSES OF UNRESTRICTED SAMPLES

Distributions of Task Performance Ratings

The contractor has made extensive analyses of the use of the 7-point scales of task performance by incumbents, peers, and supervisors. These appear in tabular form in the management report (Hahn, 1975). The following critical questions were raised:

1. Was there systematic variance of any sort in task performance ratings?
2. Were some tasks systematically rated as better performed than others?
3. Were there systematic differences between incumbent self-ratings, peer ratings of incumbents, and supervisor ratings of incumbents?
4. Were there systematic differences among the three AFSCs in the rated performance levels of tasks?

Both the contractor and AFHRL ran distributions which produced nearly identical results during the shakedown phase of the analyses. The contractor tables have been used and have been condensed to provide Figures 2 and 3. Figure 2 was derived from AIR data converted to the percentages of use of each position

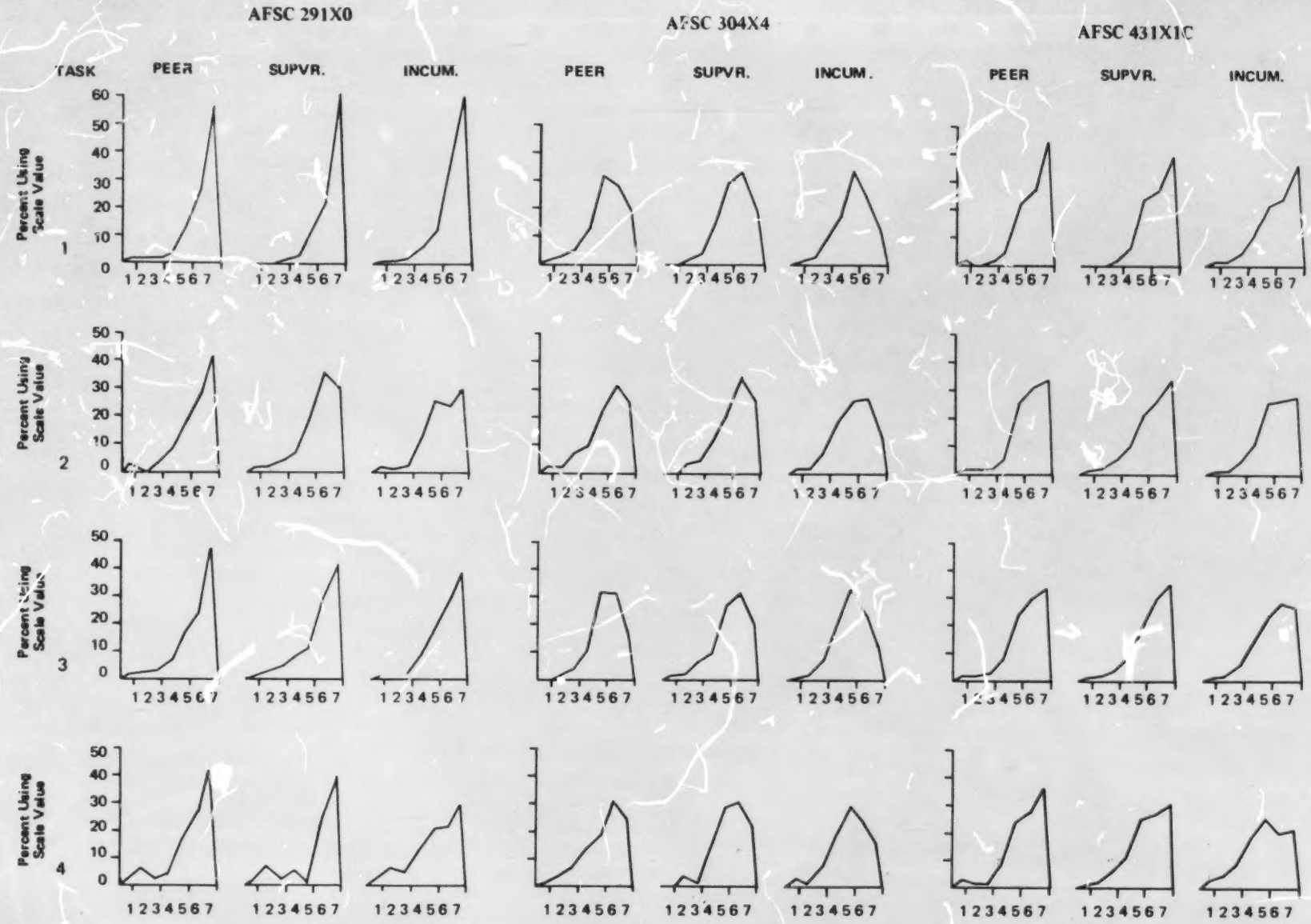

Figure 2. **Comparative use of scale values in task ratings.**

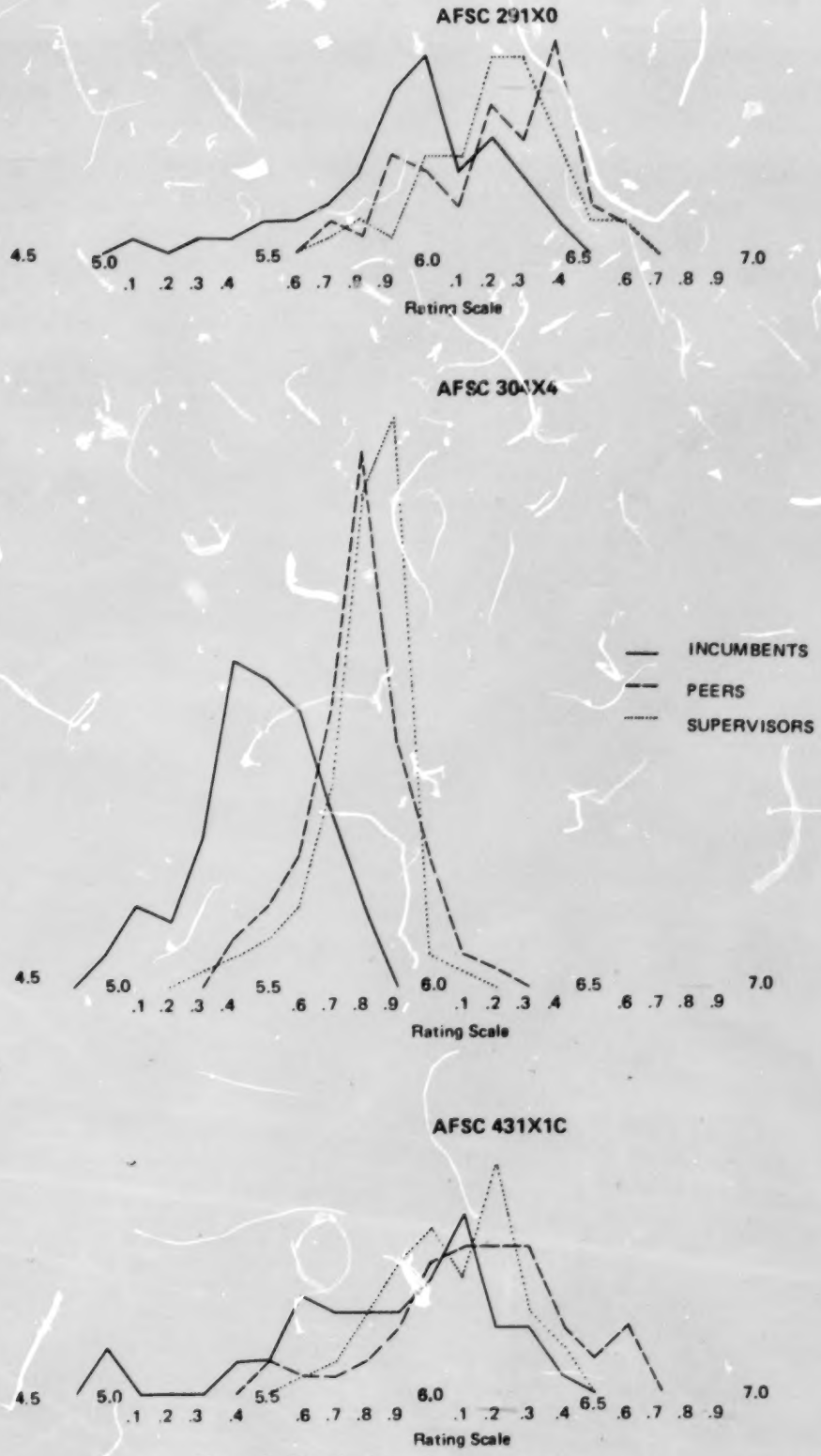

Figure 3. **Distributions of mean task ratings.**

on the 7-point scale (Tables 12 through 17 of Hahn, 1975). Only the first four entries of these tables have been shown in Figure 2 because inspection of the data revealed that these four "task distributions" were representative of the entire sets. All of the ensuing distributions are suggested in these comparison graphs. The graphs show that less use of the rating value 7 was made by personnel of the 304X4 AFSC than by the incumbents and raters of the other two AFSCs, and that most use of the 7 ratings was made by AFSC 291X0 personnel. Performance ratings were rare below the midpoint of the 7-point scale, and more likely to be self-applied by incumbents than applied to them by other raters. These graphs answer Question (1) affirmatively: There was systematic variance in the scales use.

Figure 2 was derived from Tables 1 through 6 of the management report. It is very revealing of the data and it answers several questions through distributions of mean task ratings. Since these are distributions of means whose basic n's were extremely variable, the tables were examined to determine the smallest n. It proved to be for supervisors who rated on three tasks in AFSC 304X4, and the number was 41 observations. The smallest number of incumbents rating themselves on a task was also in AFSC 304X4, and this number was 77. The distributions of Figure 3 are thus the means of task ratings based on substantial n's, some representing over 1,000 observations. It follows that the stability of these histograms is substantial. The graphs indicate that enough incumbents used the middle and the lower end of the scale to reduce the mean of the incumbent distributions markedly, which provided greater variance for incumbents. Ratings in the 304X4 AFSC are lower than those for the other two. The graphs provide these answers to the initial questions:

1. There was systematic variance in rating task performance.
2. The graphs alone could not describe the relative performance of individual tasks – no answer.
3. Some incumbents rated themselves lower than their peers and supervisors rated them.
4. There were clear differences between the task performance rating level of the three AFSCs, with AFSC 304X4 lower than the other two.

Question (2) could not be answered because the means shown in Figure 3 could have been achieved by different task compositions, without stability in the rating of any particular one. The questions could be answered by asking whether the means of ratings by incumbents, peers, and supervisors for specific tasks were converging on the same values. This was done by correlating the mean value provided in the contractor's Tables 1 through 6 (Hahn, 1975). The following results emerged:

AFSC 291X0; incumbent means versus peer means, $r_{12} = .799$
incumbent means versus supervisor means, $r_{13} = .731$
peer means versus supervisor means, $r_{23} = .899$

AFSC 304X4; incumbent means versus peer means, $r_{12} = .590$
incumbent means versus supervisor means, $r_{13} = .516$
peer means versus supervisor means, $r_{23} = .593$

AFSC 431X1C; incumbent means versus peer means, $r_{12} = .914$
incumbent means versus supervisor means, $r_{13} = .872$
peer means versus supervisor means, $r_{23} = .881$

The question of convergence is clearly answered: There were systematic differences in task difficulty for each AFSC that were recognized by all raters. Additional computations provided the evidence that ratings by peers and supervisors were converging on higher values than were those of incumbents. The grand means were as follows:

AFSC 291X0; incumbents, 5.93; peers, 6.21; supervisors, 6.20;
AFSC 304X4; incumbents, 5.47; peers, 5.80; supervisors, 5.77;
AFSC 431X1C; incumbents, 5.87; peers, 6.06; supervisors, 6.14.

The contractor has provided useful interpretations of the task performance ratings, which are quoted from the management report (Hahn, 1975):

> While there was a tendency for incumbent ratings of task level performance to pile up at the high end of the scale for all three career ladders, there were differences between ladders and there were some intra-task differences within at least two of the ladders. The reasons for such differences are not immediately apparent. The tasks within AFSC 291X0 tended to be less technical and, at least in the judgment of the contractor staff, somewhat less difficult than many in the other two AFSCs. The extreme loading on the top rating point for this AFSC could reflect a lack of actual performance variance in the field. Many of the tasks for AFSC 304X4 were of a more technical nature and tended to be more oriented toward specific hardware classes. This could account for the greater variance in obtained ratings for this AFSC. For many of the tasks somewhat explicit standards for the accuracy of task completion were available in terms of hardware tolerances. This may make it easier for incumbents to assess task performance more realistically. It should also be noted that the mean task ratings for AFSC 304X4 were lower than for the other two career fields. Even though this AFSC has only school trained input as opposed to a mixed input for AFSC 291X0, there appears to be more performance variance in the field and at least incumbent ratings of self-performance reflect this.
>
> AFSC 431X1C showed some of the extreme high end loading on many of the tasks, but there were exceptions on about one quarter of the tasks in the career field. This appears to illustrate that at least incumbents do recognize differences in performance of the various tasks which make up their jobs and are not reluctant to report such differences.
>
> The rating distributions for supervisor and peer ratings of task performance tend to reflect the same trends found in the incumbent ratings for the three career fields. . . .

The contractor has provided the standard deviations of the performance ratings by tasks whose means are shown in Figure 3. In general, the standard deviation of a set of performance ratings should be a measure of the agreement of judges, inversely, on the performance of a task. The measure is not interpreted here as a measure of agreement because of 7-point rating scale, or any rating scale, yields an ambiguous standard deviation when there is a tendency for raters to pile responses at one end of the scale.

Skill and Ability Versus Motivation Ratings

The contractor has provided frequencies converted to percentages for use of the five positions on the skill and ability versus motivation scale. These show how frequently each position was used relative to the total ratings made on a given task in Tables 18 through 20 in the management report (Hahn, 1975). Positions 1 and 2 were more frequently used on this 5-point scale than were the lower positions on the 7-point scale, which indicates that at least a number of the raters were following the intent of the instructions. Roughly, the instructions indicated that level 1 and 2 were not to be regarded as qualitatively the same as levels 3, 4, and 5. Hence, the skill and ability versus motivation scale was not strictly a scale, although it was formatted as one. In the case of raters who could not distinguish in their own minds whether a ratee's less-than-optimal task performance was due to inexperience or motivation, the scale format may have offered a chance to repeat the task performance rating without contributing new information. This is an unsettled question which is suggested by the contractor's interpretations, most of which are quoted below. It is probable that the use of the skill and ability versus motivation scale was not consistent from rater to rater. AFHRL has treated the data as though a continuous scale were being applied by correlating the ratings with those on the 7-point task performance scale for the same tasks. Coefficients around .80 appeared, which suggested that a number of the raters may have been merely repeating their ratings on a coarser scale. (However, there are indications arising from the analyses that poor motivation was perceived by raters, and that their ratings reflected it. Also, motivation was the most common criticism made in those cases where raters completed the comments page of the rating booklet.) The management report (Hahn, 1975) says:

> . . . It was based upon the concept of deficit between potential capacity and actual performance. The total scale was not truly linear. Point 5 indicated essentially no deficit. Point 4 indicated a small deficit due almost entirely to training needs. Points 3 and 2 indicated a small deficit attributable to a mixture of skill and motivational factors. Point 1 indicated a greater deficit due almost entirely to motivational factors. For some correlational analyses reported later, a somewhat risky assumption of linearity was made based on a tenuous continuum of greatest deficit to no deficit with the somewhat tacit assumption that deficits attributable to motivational factors were somewhat less desirable than those attributable to training needs. . .
>
> Review of these tables indicates that for AFSC 291X0 and AFSC 431X1C rating point 5 was the modal point for all task dimensions for supervisors and peers. . . The results for AFSC 304X4 were in

the same direction but not quite as extreme. . . .These data in conjunction with the generally lower mean task performance ratings reported previously for this career field tend to indicate that there may be a somewhat greater need for remedial action in this field than in the other two. Since the aptitude input for this field is equal to or higher than that of the other two career fields, the explanation for these slight differences must lie elsewhere. As indicated previously, the tasks for this career field tend to be highly technical and related to specific hardware items for many of which there are relatively close specified operational tolerances. Although the data do not supply a definitive answer, it is the feeling of the contractor staff that such characteristics probably account for the obtained differences.

The rating point used second most frequently in all three career fields was 4 which indicated some deficit due largely to training needs. . . .

Only a relatively small percentage of performance deficits were attributable largely to motivational factors. For AFSC 291X0 the percentages for the combined 1 and 2 categories ranged from 0 to 7 for both supervisors and peers. For AFSC 304X4 these percentages ranged from 2 to 14 for supervisors and from 2 to 20 for peers. For AFSC 431X1C the percentages ranged from 2 to 13 for supervisors and from 1 to 13 for peers.

Later, after making correlational analyses, the contractor says:

. . . From the skill and ability versus motivation ratings, it is clear that the majority of supervisors and peers feel that a large percentage of incumbents in the three fields are usually working close to their potential capacity. When deficits were reported, these were more often attributed solely to skill deficits or a combination of skill and motivation deficits than they were solely to motivational deficits or motivationally dominated mixtures with skill deficits. Slightly greater deficits were noted in AFSC 304X4 than for the other two career fields.

Correlations of Task Performance Ratings Made on the Same Incumbents

The premier question of this research is whether or not judges of incumbent's job performance will agree in their appraisals of separate task performances. If one assumes that the answer is YES, he will then need to consider several specific issues. These are outlined immediately below to identify the content of this section, but it is desirable to digress a little before presenting the answers. The issues are:

1. What is the distribution of the measures of agreement?

2. Are there systematic differences among the three AFSCs with respect to agreement on task performance ratings?

3. Is one kind of rater measurably superior to another, such as a peer being a better judge than a supervisor?

4. Do incumbents contribute usable data on self-ratings of task performance?

5. Do the data yield any clues that will differentiate tasks with respect to the ratability of their performance?

Some of these points were covered in the course of preliminary analyses and the decision was made not to pursue them further. The direction which the analyses took was partly determined by computations undertaken by AFHRL on the same data bank as used by AIR, and concomitantly. These produced the discovery that certain approaches would not be likely to be productive. They are mentioned here so that other investigators can avoid efforts which might prove to be wasted. It should be kept in mind that when these analyses were begun, there existed no guidelines as to what one should expect to find in operational ratings of task performance.

It is normally assumed that psychological measures converge on stable values when more observations are added to their computed mean. With this concept, it would be assumed that if one correlates the task performance ratings made on an incumbent by a peer and a supervisor, then goes to the data bank to find a third rater, he could expect the third rater to stabilize the data. He would expect that if he computes the mean of the peer and supervisor ratings, then correlates the new rater with (a) the peer, (b) the supervisor, and (c) the mean of the peer and supervisor's ratings, the third correlation would be the highest. This was done and it was continued out to four raters where available. While all the correlation coefficients were

rather unstable, correlations with means of two raters were slightly higher than correlations between single raters. Now a new mean of three ratings was available, and all possible combinations were computed. The net result was that there was no practical increase in the set of correlation coefficients beyond computing the mean of two raters, almost any two who rated a single task performance of an incumbent. This could have been a function of special operational situations, since only certain types of tasks could be performed by enough people to provide several observers of the same incumbent. Bearing in mind that additional raters must be equally able to observe the incumbent's work, the failure to increase reliability may have been a function of adding random variance to an unstable set of data, where to achieve a stable value would require more raters than could exist. From the experimenter's standpoint: in this case two raters were sufficient and procuring additional raters might not be cost effective.

As one adds raters to stabilize the task performance rating mean of a specific incumbent on a specific task, he is not doing the same thing as was done to produce the data of Figure 3. The means of task ratings for all raters performing a given task appear to behave as do most accepted statistics, converging on stable values. In the previous correlations between task means obtained from incumbents, peers, and supervisors for AFSC 304X4 it was found that the coefficients were lower than those for the other two AFSCs. It seems probable that the reason was the fact that raters in AFSC 304X4 used the 7 value of the scale less than the raters in the other two AFSCs, which left instability in the upper end of the scale as well as the lower. Put another way, there was room for individuals to vary at the top of the AFSC 304X4 task performance. We shall see that this may have contributed to making the performance of these personnel much more measurable than was that of AFSCs 291X0 and 431X1C.

It turned out from these early analyses that incumbent self-ratings correlated less well with ratings assigned to them by either peers or supervisors than the peer and supervisor's ratings correlated with each other. It was evident that the most fruitful data would come from observers who were objective rather than subjective. One should note that the matter of objectivity was primarily the role that an individual was playing, since an incumbent became a peer or a supervisor when he was rating someone else.

The contractor pursued the problem of peer and supervisor agreement in detail, reporting on task performance ratings and the skill and ability versus motivation ratings in Tables 22 through 24 of the management report (Hahn, 1975). He gives the number of incumbents who rated on each task along with the peer and supervisor correlations for both the task performance and the skill and ability versus motivation scales. Accompanying these is the probability estimate that an obtained coefficient could have occurred by chance based upon the likelihood that in 1,000 trials the coefficient could have been a deviation from a true r of zero. This is a strict application of the probability assumptions because it treats each coefficient as though it were independently obtained. The probability that the whole distribution would have varied from zero in the positive direction is not considered.

The skill and ability versus motivation ratings will be disposed of first by dismissing them. These ratings probably had validity for a number of raters, but the distribution of correlations suggests that they were a less reliable reflection of the task performance ratings which immediately preceded them. For the most part, the coefficients in the skill and ability versus motivation column are smaller than their corresponding task performance rating correlations, and they follow the same pattern. It should not be inferred from this that the concept of rating skill and ability versus motivation was invalid; the difficulty lay in a combination of format and proximity to a rating just made. The concept should still be tested.

The correlations provided by the contractor answer the basic question of how well raters agreed on evaluating task performance. These have been reduced to graphic form in Figure 4. The illustration is three sets of bar graphs, the upper tier representing the coefficients obtained when less than 100 incumbents were rated on the performance of a task, the middle tier representing coefficients for 100 to 199 incumbents, and the lowest tier representing correlation coefficients for 200 or more ratees. The baseline places correlation coefficients in intervals of .05 and the frequencies are shown in units of 1. Bar graphs appear for each AFSC within an interval in the following order: 291X0, 304X4, and 431X1C. A dashed line has been drawn through one of the intervals in each tier, specifically, $r = .325$, $r = .225$, and $r = .175$. These are

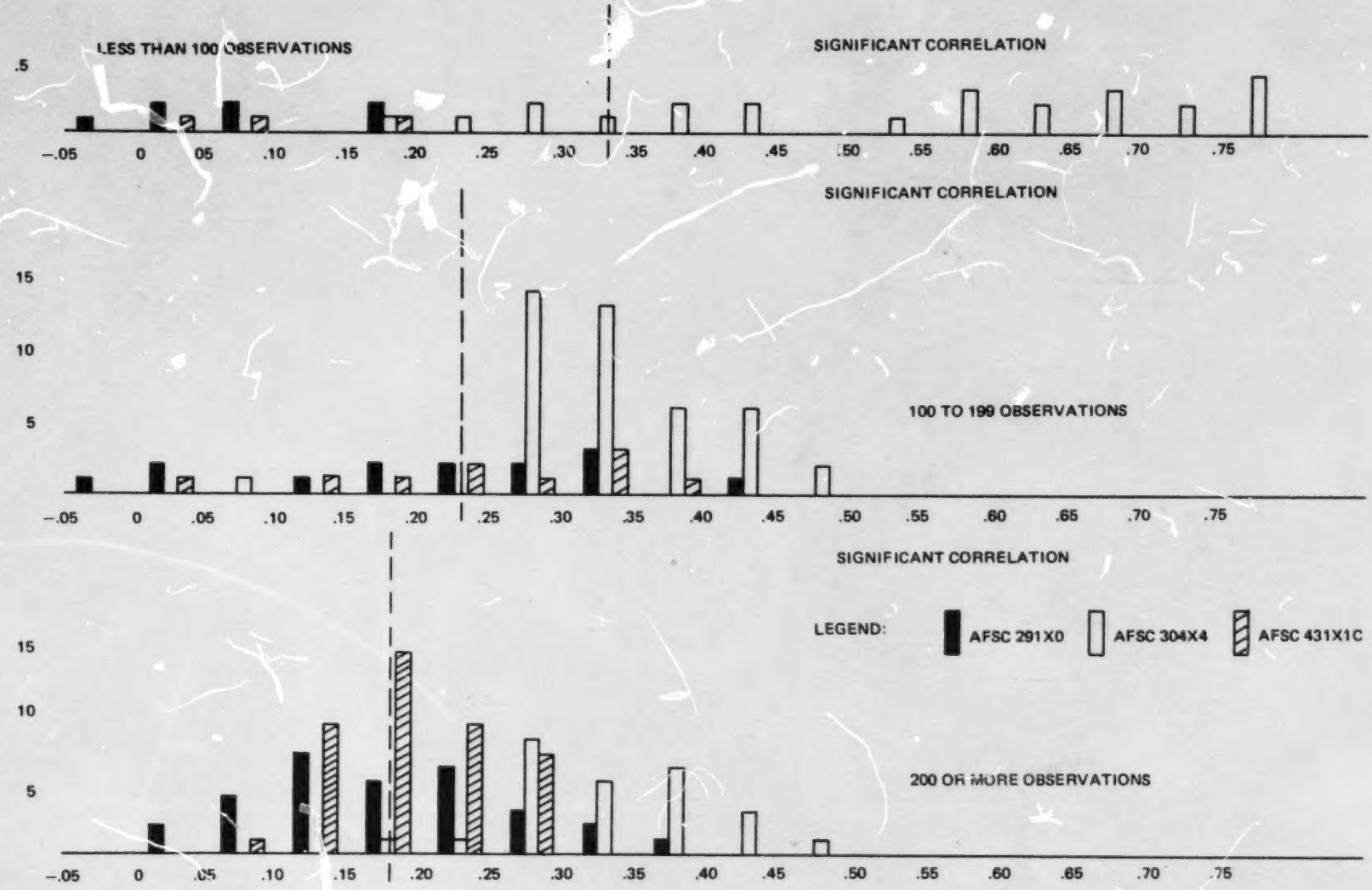
LESS THAN 100 OBSERVATIONS
SIGNIFICANT CORRELATION
.5
−.05 0 .05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75
SIGNIFICANT CORRELATION
15
10
5
100 TO 199 OBSERVATIONS
−.05 0 .05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75
SIGNIFICANT CORRELATION
LEGEND:
AFSC 291X0
AFSC 304X4
AFSC 431X1C
15
10
5
200 OR MORE OBSERVATIONS
−.05 0 .05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75

Figure 4. **Peer vs. supervisor task performance rating correlations.**

approximate values for correlations different from zero at the 01 confidence level, based respectively upon 50 observations, 150 observations, and 250 observations per correlation. The bar graphs represent the peer versus supervisor correlations in the task performance rating columns of Tables 22 through 24 of the management report (Hahn, 1975). Although the dashed lines are a rough approximation, by actual count of the contractor's r's with computed significance at the .01 level or better there is very little difference between the number to the right of the dashed line and the number of significant r's which the tables provide. The number of coefficients in Figure 4 depends upon the number of tasks rated, which was 51 for AFSC 291X0, 92 for AFSC 304X4, and 55 for AFSC 431X1C.

Returning to the questions at issue – (1) What is the range and frequency (distribution) of the measures of agreement? –it is seen that the size of the correlations is a function of the AFSC. As a set of statistics, the correlations for AFSC 304X4 could not have arisen from the same population as those of the other two AFSCs within any reasonable probability. Only 5 of the 92 AFSC 304X4 coefficients are nonsignificant, 16 of the 55 coefficients for AFSC 431X1C are nonsignificant, and 33 of the 51 coefficients for AFSC 291X0 are nonsignificant. These data answer question (2). Are there systematic differences between the AFSCs with respect to agreement on task performance ratings? Collapsing the three tiers into one distribution for each AFSC, one finds that the median r for AFSC 304X4 lies in the interval .30–.35 and that the medians for both AFSC 291X0 and 431X1C lie in the interval .15–.20. Without qualification, AFSC 304X4 is superior to the other two specialties in task performance rating agreement, and AFSC 431X1C appears to have had more ratable tasks than AFSC 291X0.

This section of the analyses was begun with the assumption that raters could agree upon task performance evaluations. Had they not been able to agree to some extent, the bars of Figure 4 would have clustered about the value zero.

The last three questions raised in the analyses of rater agreement will be given tentative answers without data to support them at this point. Question (3) – is one kind of rater measurably superior to another? – was answered partially by the preliminary analyses performed by AFHRL when it was found that peer ratings correlated as well with pooled ratings as did supervisor's ratings but that incumbents self-ratings yielded lower correlations. Additional data were amassed when computing other relationships, which suggested that in most respects supervisors were more reliable sources of judgments than were peers. By that time analyses of incumbent data had been dropped. Incumbents could not be correlated with incumbents as peers could be with peers and supervisors with supervisors. If the incumbent was providing valid unique data, that fact would have to be reached in a different manner. Thus, question (4) – Do incumbents contribute usable data on self-ratings of task performance? – has not actually been answered, and it has not been eliminated from future consideration. Finally, question (5) – Do the data yield any clues that will differentiate tasks with respect to the ratability of their performance? – was noted by the contractor. This question is the essence of the research problem and it will be dealt with later.

Prediction of Task Performance Ratings from Data of Record and Cognitive Test Scores

The justification for the entire enterprise is to show that rating task performance contributes to understanding whole job performance. This requires that task performance ratings be predictable or accountable from something. That is, do tasks in one specialty reflect aptitude, while tasks in another specialty reflect other factors, such as length of service, social skills, attitudes, grade, or some aptitude as yet unmeasured? Or do both sets of task ratings show no predictability from anything that is consistently measurable?

Obviously, one must dispense with this last possibility before he needs to trouble himself about the others. The contractor has provided relevant analyses in a predictor-by-predictor discussion in the management report (Hahn, 1975). These analyses have been summarized in Table 2; it may be well to point out how the data have been assembled to produce its cell entries, for it is the product of a number of steps. A great deal of data were collected in the surveys, and these data were augmented by giving the contractor

Table 2. **Significant Correlations with Task Performance Ratings**

Predictor Variable	AFSC 291X0 (Base 51 tasks)			AFSC 304X4 (Base 92 tasks)			AFSC 431X1C (Base 55 tasks)		
	Peer	Supvr.	Comp.	Peer	Supvr.	Comp.	Peer	Supvr.	Comp.
Grade	15	13	18	61	43	52	22	35	20
Months on base	13	25	27	2	14	10	2	13	8
Months in AFSC	9	9	8	58	3	29	4	19	6
TAFMS[a]	6	9	8	57	9	47	6	11	5
Decoding score	3	0	0	0	6	3	2	3	3
Memory for landmarks	7	5	4	6	24	3	4	1	2
Complex scale reading	7	1	9	3	2	0	3	1	11
Pursuit	3	0	3	1	1	1	1	9	11
Figure analogies	9	5	8	3	18	2	1	9	1
Hands	3	3	0	1	8	4	3	3	1
Cubes	1	1	0	5	5	10	2	0	6
Mechanical principles	0	0	0	15	42	32	5	2	5
Following directions	2	0	1	4	13	3	1	1	2
Practical estimations	8	9	3	0	16	9	3	2	3
Spatial reasoning	1	1	1	7	2	9	2	3	3
Marital status	6	9	10	6	6	4	0	1	1
Size of city of origin	1	3	4	2	2	2	1	15	4
Mechanical AI	3	11	1	10	2	3	15	5	9
Administrative AI	1	24	3	3	2	1	0	2	2
General AI	0	5	1	3	1	2	2	1	2
Electronic AI	1	8	0	8	12	15	2	0	0
Sex	2	6	6	—	—	—	—	—	—
Year of enlistment	5	12	12	50	15	24	8	45	10
Education level of enlistment	2	4	0	71	23	57	20	35	14
AFQT score	2	7	2	4	1	1	3	1	1
Mean	4.4	6.8	5.2	15.2	10.8	12.9	4.5	8.7	5.2
Mean in %	8.6	13.3	10.2	16.5	11.7	14.0	8.2	15.8	9.5

[a]TAFMS = total active federal military service.

information obtained from Air Force record tapes maintained by AFHRL. Incumbent names and Social Security numbers were matched with the tape records. Rather than enumerate all the items collected from official records, the ones that were differently obtained are cited. The 11 cognitive tests were part of the surveys. The months on base and months in AFSC were taken from the job inventory information supplied by the incumbent. Months on base cannot be extracted accurately from the official records files because leave time and other inconsistencies make real time on base and official time different. Cumulative items, such as time in AFSC, are better given by the incumbent, for similar reasons. Using predictor variables which employed job inventory data automatically excluded incumbents sampled in the supplemental survey, since that survey did not administer job inventories. The use of cognitive test score predictors meant that any personnel who failed to take the battery were excluded. Lack of experimental test score data was the largest single source of data attrition in the entire study. To round out the list of selective effects of predictor variables, the size of city of origin was an item in the Biographical Inventory, one of the noncognitive instruments of the experimental battery. The remaining predictors listed in Table 2 came either from official record tapes or were corroborated by the records. These were quite complete and did not represent a factor which could select incumbent data.

Thus, from predictor to predictor there were conditions which affected the sampling of ratees in each correlation coefficient. This does not cover the sampling of raters, where such factors as shift work and the

kind of task performed operated to select raters. In computing a correlation coefficient the contractor appears to have tried to maximize the number of observations in the criterion vector. This was legitimate and desirable, but it must be appreciated in order to interpret the data of Table 2. The number of peers or supervisors rating an incumbent on a specific task was so variable that it is inconceivable that any two coefficients would have depended upon precisely the same ratees and raters. Thus, the cell entries of Table 2 must represent counts of correlations with extremely variable numbers of observations. These coefficients were small, though usually positive, probably ranging from −.17 to .27, but having fairly substantial n's. They could consequently be identified as significantly different from a true coefficient of zero, and this was reported at the 5 percent level of confidence. The entries of Table 2 are a simple count of occurrence for each cell. These depend upon the number of tasks rated in each specialty, 51 for AFSC 291X0, 92 for AFSC 304X4, and 55 for AFSC 431X1C. For all practical purposes the count of cells for AFSC 291X0 can be compared directly to the count for AFSC 431X1C; and to provide a rough approximation, the equivalent count for AFSC 304X4 can be obtained by dividing the cell entry by 2.

As the contractor has pointed out in his summary of these analyses, 1 in 20 of the coefficients could appear as significant from chance occurrences. The table reveals that some predictors yielded frequencies far above chance for all three specialties, grade and year of enlistment being notable producers of nonchance coefficients. Both of these predictors reflect time in service, and year of enlistment also reflects the chronological age of the incumbent. As one looks at the other predictor variables, it becomes difficult to determine which are predictive, or, if any really were. However, the table provides an answer to the most serious doubt raised; some predictors *did* make statistically significant predictions of task performance ratings. Furthermore, enough nonchance predictions occurred to indicate that several variables were correlated with task performance beyond chance. The mean number of nonchance correlations was computed for each column, and in the bottom row it is shown as a percent. The values obtained exceed the 5 percent limit, and even when grade is removed and the mean is recomputed, the percent values exceed chance expectation.

The prediction of the composite ratings shown in the third column lies between that of the low and high group for each AFSC. While this would seem to be a simple arithmetic necessity, in fact it is not. The composites are made up of two raters, which eliminates unpaired cases from the two columns preceding and greatly reduces the number of observations. This, in turn, lowers the probability of a significant correlation coefficient appearing.

In the 291X0 and 431X1C AFSCs the larger number of nonchance correlations arose from supervisor ratings, and in the 304X4 AFSC the peer ratings produced more nonchance r's. A similar observation was found in subsequent analyses, suggesting that some factor (such as opportunity to observe the ratee) may have operated to differentiate the AFSCs.

At best, the data of Table 2 indicate trends. The interdependency of such variables as length of service and grade require that regression approaches be used to determine their unique value as performance predictors. The analyses were performed and are presented later.

Correlations Among the Six Overall Performance Rating Dimensions

Six overall performance ratings were made by peers and supervisors upon completion of the task dimension ratings. Since the number of task dimensions was different for each AFSC, and represented an appreciable reading time, the number of ratings made before the six final ratings differed considerably for the three specialties. Task ratings were successive acts, involving marks on pages of the survey booklet and sufficient reading of each for the rater to decide whether or not he would rate performance on a particular task. The 92 (actually 95) task dimensions of AFSC 304X4 represented nearly twice the reading time of the 51 task dimensions of AFSC 291X0. On the other hand, the six overall performance rating dimensions all appeared on the same page of the booklet, and their reading time was relatively brief. Consequently, one would expect that, if the raters were performing in a perfunctory way, the correlations among the six scales would be very high. They are shown in Table 3 to have been high, but not as inflated as one might expect

Table 3. **Correlations Among Six Overall Performance Rating Dimensions**

AFSC	Dimension	Ratings By Peers					Ratings By Supervisors				
		GP	Quan	Qual	ES	SI	SK	GP	Quan	Qual	ES	SI	SK
		554–632 incumbents rated						796–853 incumbents rated					
291X0	General performance (GP)		.82	.85	.78	.81	.74		.87	.87	.79	.81	.78
	Quantity of work (Quan)			.80	.80	.81	.71			.86	.80	.82	.77
	Quality of work (Qual)				.76	.78	.72				.74	.77	.74
	Exceeds his share (ES)					.88	.76					.92	.78
	Self-initiating (SI)						.79						.82
	Shares knowledge (SK)												
		371–419 incumbents rated						420–447 incumbents rated					
304X4	General performance (GP)		.82	.85	.74	.78	.72		.89	.86	.81	.83	.80
	Quantity of work (Quan)			.72	.78	.75	.64			.83	.84	.84	.76
	Quality of work (Qual)				.67	.71	.68				.76	.80	.76
	Exceeds his share (ES)					.85	.66					.89	.77
	Self-initiating (SI)						.68						.80
	Shares knowledge (SK)												
		539–619 incumbents rated						747–796 incumbents rated					
431X1C	General performance (GP)		.87	.85	.79	.80	.76		.86	.87	.80	.84	.80
	Quantity of work (Quan)			.79	.82	.78	.75			.84	.84	.83	.80
	Quality of work (Qual)				.75	.77	.73				.79	.81	.78
	Exceeds his share (ES)					.87	.73					.91	.83
	Self-initiating (SI)						.77						.85
	Shares knowledge (SK)												

Note: — Number of ratings for correlation matrices are:
a) 291X0 peer = 554-632 incumbents
b) 291X0 supv = 796-853 incumbents
c) 304X4 peer = 371-419 incumbents
d) 304X4 supv = 420-447 incumbents
e) 431X1C peer = 539-619 incumbents
f) 431X1C supv = 747-796 incumbents

from repetition of a single overall impression. Table 3 is not strictly an intercorrelation matrix. It is a combination of the contractor's Tables 28 and 29 in the management report (Hahn, 1975). The n's of the cells of Table 3 vary slightly, and different incumbents are represented among the ratees by peers and supervisors. However, the basic sample underlying these correlation coefficients is a large unit, and the values shown are quite similar to those found by AFHRL for samples reduced to the same incumbents. The coefficients are not so large as to preclude interpretable differences among the dimensions of overall performance.

During preliminary analyses of these data AFHRL segregated samples containing all six ratings by peers and supervisors on the same incumbents. The samples ranged from 300 to 500 ratees per AFSC. Using regression techniques in which it was assumed that the general overall rating was the criterion, the other five ratings were made to account for the general performance rating. The intercorrelation matrix produced correlations in the low .80's when the rater's own scales were correlated, which is very similar to what appears in Table 3. When peers were correlated with supervisors, the r's were in the mid- .30's. Employing the peers' general overall rating as criterion, the supervisor's five subscale ratings were used as predictors, and then the supervisor's general rating was used as criterion and the peers' subscale ratings were the predictors. The results were suggestive of untapped relationships among the six overall scales. The regressions indicated that the three specialties might be dissimilar with respect to the subscales and their relative weights in accounting for overall or general performance assessments.

In sum, the overall ratings were highly related to each other when taken from the same page of a rating booklet. Part of this correlation would necessarily be attributable to rater tendency to use the top of the 100-point overall rating scale, or not to use it. Removing the rater tendency factor greatly reduced the correlations, and the regression problems provided suggestions that there were complex relationships among the six ratings which reflected differences among the specialties. This was an important lead in terms of its effect upon the course of subsequent analyses.

Correlations Between Task Performance Ratings and Overall Performance Dimension Ratings

The contractor has provided tables of correlations between task performance ratings and the six overall dimensions of performance for both peers and supervisors. He has included identification of those correlations which fail to be significant at the .01 level of confidence. They are Tables 30 through 35 in the management report (Hahn, 1975). These have been reduced to graphic form, and they appear as Figures 5 and 6. Figure 5 contains the general overall rating, the quantity of work rating, and the quality of work rating. The three ratings in Figure 6 can be regarded as motivational. Correlations between peer estimates are shown as solid lines and correlations between supervisor estimates as broken lines. The graphs indicate that in all respects of agreement between task ratings and overall ratings personnel in the 304X4 AFSC provided higher correlations than did those in the other two specialties, with modal r's 10 points higher than those of the other two specialties. AFSC 291X0 yielded the lowest agreement, though this was still in the modal range of r = .50. A second feature of these graphs is that, with a lone exception, the correlations obtained from supervisors were higher than those obtained from peers. The exception occurs in the quantity of work performed in the 291X0 AFSC. It is interesting because it is one of the few instances for which an explanation can be hazarded. The hypothesis is that this difference is genuine and reflects a difference of opportunity between peers and supervisors to observe incumbents in the 291X0 specialty activities. Continuous records are maintained in a communications center of messages sent and received, of circuit usage, overloading, etc. Peers tend to move together on shifts, which do not correspond as closely to the assignment of supervisors. This provides a peer with a better estimate of the incumbent's normal production. If the explanation should prove to be valid, it would attest to the sensitivity of the methodology.

Summary of Findings for Unrestricted Samples

The contractor made an effort to maximize the size of the samples, and for that reason the products of the contractor's analyses have been given precedence for presentation over parallel analyses performed by AFHRL. At this point and hereafter, the analyses reported will be those performed by AFHRL. Preparation of samples for regression analyses required selection of data so that all required variables would present complete data for intercorrelation purposes. This greatly limited the cases that could be used because there were many concomitant variables for each incumbent. Thus, the foregoing analyses could better reflect the variables from which the data of the regression problems were taken, and it becomes the responsibility of the regression analyst to show that his samples were representative of the full population of the study. The findings up to this point show that:

1. Task performance in each AFSC could be rated with greater than chance agreement between raters.

2. The specialty, AFSC 304X4, which the contractor considered to have the best established criteria of task performance in an objective sense, yielded the least use of the top of the rating scale and the highest correlations between raters on all performance estimates. It also yielded the highest within-rater correlations.

3. AFSC 291X0, with tasks which the contractor considered likely to be the easiest to learn, yielded the greatest use of the top of the performance rating scale and the lowest correlations between raters.

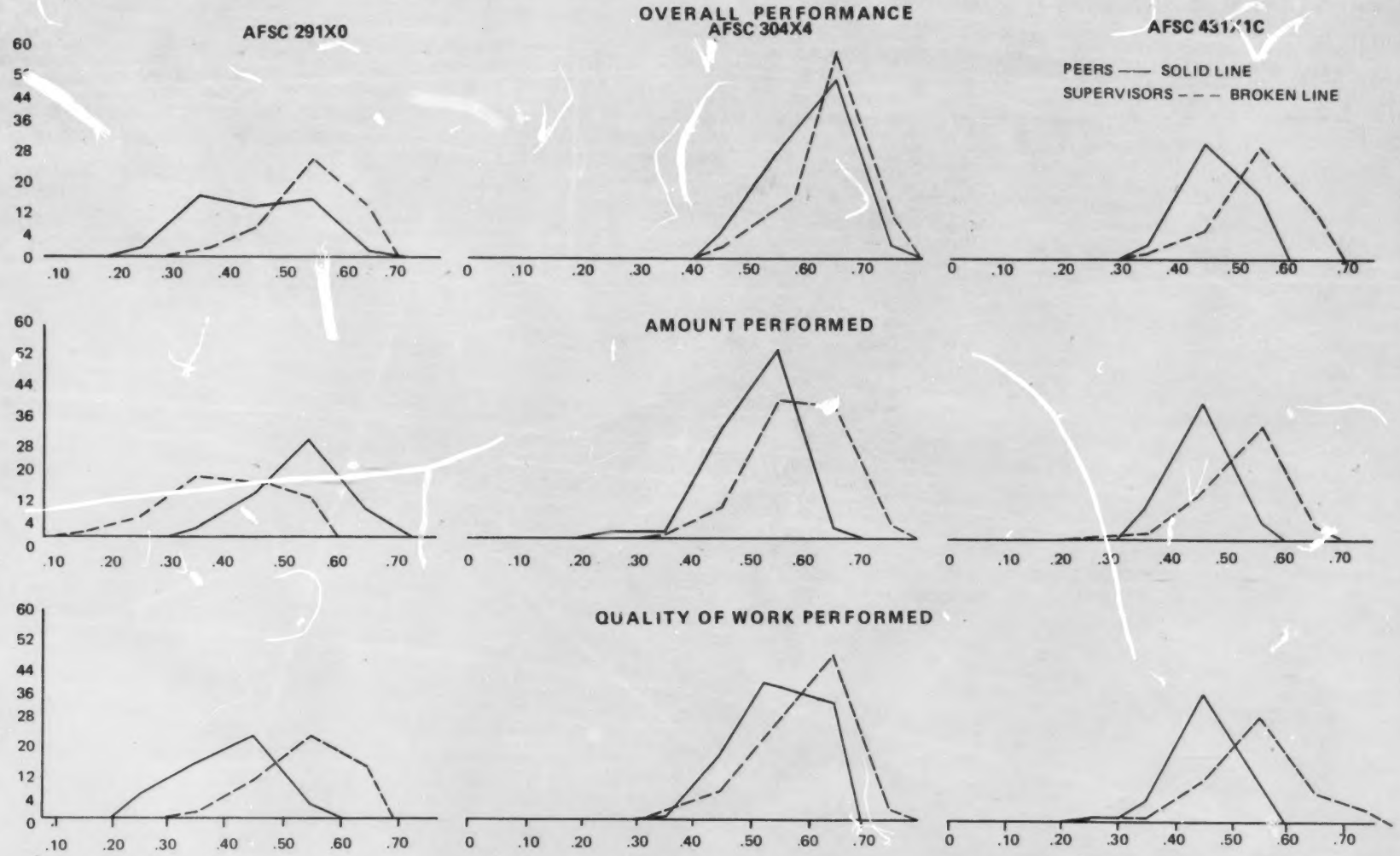

Figure 5. Correlations of Task and General Performance Ratings, Product.

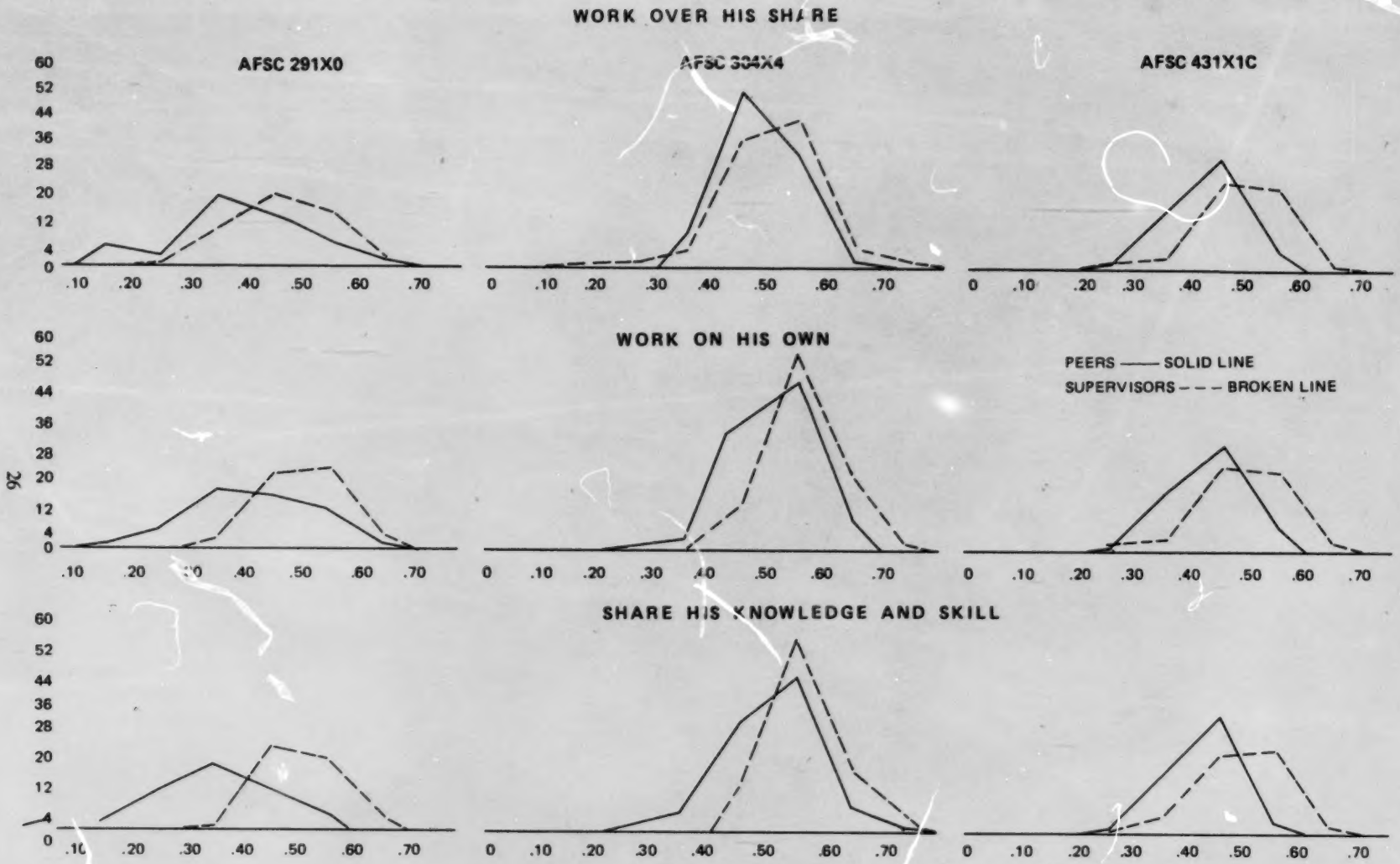

Figure 6. Correlations of Task and General Performance Ratings, Motivation.

4. Incumbents frequently rated themselves lower on task performance than they were rated by their peers or supervisors. The uncertainties raised by their data put their ratings in doubt as a source of useful performance estimates.

5. There were systematic differences within specialties among task performance ratings, as well as clear differences in mean task performance ratings among the three specialties.

6. The AFSC whose incumbents had the highest aptitude, 304X4, yielded the lowest task and overall performance means.

7. Despite the fact that shift operation may have reduced the supervisors' opportunity to observe incumbent performance below that of peers, supervisor ratings of task performance appeared to be slightly more reliable than peer ratings.

8. The number of tasks rated prior to making overall performance ratings had no bearing on the correlation between task and performance ratings. AFSC 304X4, with far the most tasks, yielded the highest correlations between individual task ratings and overall performance ratings.

9. Correlations between data of record, such as grade or aptitude test scores, and task performance ratings tended to be significantly greater than zero. Although the r's were small, they were systematically arrayed instead of randomly distributed.

VIII. REGRESSION ANALYSES OF RESTRICTED SAMPLES

The maximum sample analyses made up to this point have been representative of the data pool. However, use of the incumbent records with incomplete data, such as test results, has raised questions regarding sample selection for future analyses to be made on variables that have not been reported here. These variables are primarily in the areas of vocational interest and job satisfaction. The monitoring agency, AFHRL, has performed all the ensuing regression problems in order to select a "flagged" set of rating data which would meet requirements for future analyses.

The "Flagged" Sample Concept

A basic population of criterion ratings was so chosen that any new set of analyses would involve identifiable subsets of the parent population, not replacements by samples of unknown representativeness. This rule will hold for all studies in which a complete set of criterion ratings is required, including studies not contained in the present report.

The first condition for selecting a data member was that all six of the overall performance ratings were present. Supervisors usually assigned all six ratings, but peers often skipped the first (general performance) although going on to rate on the five subdimensions. Such a record was not used, and another peer or supervisor record containing full data was chosen. The goal was to be able to predict every kind of performance rating for the entire population. Future regression problems involving new predictor variables would necessarily encounter incumbents with missing data in other than overall performance measures. Selecting a flagged population makes it possible to rerun any of the original analyses on the subsample to determine if selection on a new variable has biased the sample. The criteria for selecting a record to be used in the flagged population were:

1. All six overall performance ratings were available;
2. performance was rated on at least one task;
3. a combination of one peer and one supervisor was preferred, but not required;
4. where there was a choice between two peers or two supervisors, the rater who had rated on more tasks was chosen;
5. not more than two records per incumbent were used;
6. in the absence of two rating records one record was used if it was complete. (This was most often a record from the supplemental survey, which employed one rater, a supervisor.)

Combining Raters

A serious difficulty arises when many of the ratees perform different tasks. One must have a true intercorrelation matrix to compute regression problems: a partial overlap will not suffice. Nevertheless, it is possible to predict overall performance ratings under these conditions by generating a matrix containing a membership variable (dummy variable) for each task. The resulting intercorrelation matrix contains the criterion vectors (six in this case) and 2N predictor variables, where N is the number of rated tasks. The task rating data cells of the computation matrix contain zeros for missing data and numerical values where ratings were assigned. The dummy vectors contain 1s where the corresponding rating vectors contain 0s and 0s where the corresponding vectors contain ratings. If the prediction system also contains other types of data, such as grade and months of service, it is possible to handle missing data in the same manner. However, it was not necessary to do so in these analyses because relatively few incumbents were missing demographic data of importance. Loss of experimental test battery data was regarded as a basis for dropping the case, but regression problems were first computed for the entire flagged population in order to determine if the samples which had been retained because they had test data were biased.

The principle of one incumbent, one criterion rating, was applied. If there were two criterion raters, the mean rating was computed, or if there was one criterion rater, his rating was used with weight equal to that of the two. Several combinations could exist for the task performance ratings, however, An incumbent with two criterion raters might be rated on the performance of a task by one or both of them. It was necessary to decide whether to use the task ratings separately or to compute their mean as a single predictor. It might be argued that using the task ratings separately would afford a more defensible prediction system, since it takes account of those raters who failed to observe performance on a given task. Regression problems were run with the six criteria for each AFSC using the two modes, the mean task performance rating and the two ratings separately. The resulting R^2s were very close, with the greatest observed difference .017. In the 18 possible R^2s (six criteria for each of three AFSCs) 15 computed with mean predictors were larger. It was accordingly decided to use a one-man, one-set of data, analysis. This greatly simplified the problem when data of record were introduced into the prediction system. The R^2s presented in Table 4, and following, were so computed. Table 4 is arranged in three major columns by

Table 4. **Task Performance Ratings Predicting Overall Performance**
(Flagged Population)

Criterion	AFSC 291X0		AFSC 304X4		AFSC 431X1C	
	R^2	R	R^2	R	R^2	R
General performance	.2093	.457	.2905	.539	.2255	.475
Amount of work performed	.2241	.473	.3066	.554	.2089	.457
Quality of work	.2321	.482	.2980	.546	.2042	.452
Will exceed his share	.2244	.474	.2935	.542	.2265	.476
Self-initiating	.2172	.466	.2835	.532	.2242	.474
Will share his knowledge and skill	.2291	.479	.2802	.529	.2089	.457
Rating Characteristics						
Number of tasks	51		95		55	
Number of predictors	102		190		110	
Number of incumbents	1,254		806		1,195	

AFSC. The actual statistics fall under the headings R^2 and R. The last three rows provide the number of tasks which is doubled by adding the dummy variables in the next row, and the number of job incumbents in the final row. Thus, AFSC 291X0 had 51 rateable tasks, which produced a prediction system of 102 predictors when accompanied by the dummy variables. There were 1,254 incumbents in AFSC 291X0, and this meant that each correlation coefficient in the regression equation matrix contained 1,254 observations. The tasks of AFSC 291X0, together with their nonrated membership variables, yielded an R^2 of .2093 with the criterion of general overall performance, accounting for about 21 percent of the criterion rating variance. The table contains an R^2 range from .2042 to .3066, and AFSC 304X4 is highest on each of the six criteria.

Prediction from Data of Record

Tables 5, 6, and 7 give the zero-order correlations for the 25 demographic and aptitude variables used in subsequent regression problems to predict the six-criteria overall performance ratings. These variables include most of those used by the contractor and previously identified in this report. The incumbent's age on entering the Air Force has been employed instead of his date of entry into service. Unfortunately the variables of time on base and time in AFSC had to be dropped. These were only obtainable from job inventory responses, which were unavailable for ratees surveyed in the supplementary sample. In subsequent analyses the 25 variables have been classed as demographic and aptitude. The aptitude variables consist of 11 experimental tests and the four Aptitude Indexes; the remaining 10 variables are the demographic group.

Table 5. **Correlations with Performance Ratings AFSC 291X0**
N = 457

Predictor Variable	Criterion[a]					
	General 1	Quantity 2	Quality 3	Exceeds Share 4	Self Motivating 5	Shares Knowledge 6
Grade	1058	1449	0969	0864	1029	1068
Total active federal military service, TAFMS	0661	0950	0587	0814	0934	0806
Decoding test score	0361	0162	0459	0249	0381	0350
Memory for landmarks test score	1008	0688	1208	0623	0857	0820
Complex scale reading test score	0200	−0099	0222	−0357	−0268	0024
Pursuit test score	0121	0476	0423	0282	0518	0533
Figure analogies test score	0301	0231	0407	0329	0199	0614
Hands test score	0985	0840	0673	0823	0728	0479
Cubes test score	−0385	−0401	0047	−0057	−0111	−0139
Mechanical principles test score	−0420	−0524	−0626	−0649	−0623	−0318
Following directions test score	0664	0098	0560	−0157	0082	0484
Practical estimations test score	0460	0323	0725	−0358	0022	0204
Spatial reasoning test score	1053	0630	0994	0649	0450	0640
Coded 1 if married/ 0 if otherwise or unknown	−0131	0142	−0304	−0450	−0426	−0226
Coded 1 if divorced or separated/ 0 if otherwise, unk.	0747	0702	0591	0628	0411	0591
Code 1 if single/ 0 if otherwise or unknown	0058	−0203	0260	0409	0444	0201
Size of city of origin	−0401	−0128	−0300	−0028	−0183	−0145
Mechanical Aptitude Index	−0112	−0317	0219	−0463	−0077	−0060
Administrative Aptitude Index	−0040	−0269	0036	−0226	−0321	−0083
General Aptitude Index	0078	−0100	0334	−0180	−0274	−0006
Electronic Aptitude Index	−0409	−0683	−0521	−0451	−0624	−0162
Coded 1 if male/ 0 if female or unknown	−0697	−0623	−0946	−1171	−0896	−0773
Coded 1 if female/ 0 if male or unknown	0697	0623	0946	1171	0896	0773
Age at time of enlistment	0536	0867	0504	0849	0977	0798
Education level at time of enlistment	0036	0010	−0157	0187	0309	0241

[a]Decimal points have been omitted.

Table 6. **Correlations with Performance Ratings AFSC 304X4**

N = 399

Predictor Variable	Criterion[a] General 1	Quantity 2	Quality 3	Exceeds Share 4	Self Motivating 5	Shares Knowledge 6
Grade	1357	1020	1220	0793	1041	1234
Total active federal military service, TAFMS	0913	0670	0912	0501	0830	0546
Decoding test score	0817	0270	0700	0406	0300	0627
Memory for landmarks test score	0748	0291	0686	0888	0412	0611
Complex scale reading test score	1243	0656	1185	1071	0810	1216
Pursuit test score	0641	0212	0544	0399	0261	0548
Figure analogies test score	−0449	−0503	−0189	0053	−0530	0160
Hands test score	0435	0161	0306	0838	0583	0664
Cubes test score	0187	−0110	−0006	0351	0498	0022
Mechanical principles test score	0362	0304	0524	0704	0360	0811
Following directions test score	0075	−0195	0372	−0194	−0285	0347
Practical estimations test score	−0240	−0171	−0161	0202	0022	0053
Spatial reasoning test score	0598	0059	0637	0388	0104	0710
Coded 1 if married/ 0 if otherwise or unknown	0368	0085	0205	0067	0162	0292
Coded 1 if divorced or separated/ 0 if otherwise, unk.	0603	0590	0456	0531	0496	0304
Coded 1 if single/ 0 if otherwise or unknown	−0492	−0205	−0298	−0175	−0264	−0355
Size of city of origin	0240	0329	−0041	0142	0155	0157
Mechanical Aptitude Index	0461	−0365	0103	−0046	0130	0012
Administrative Aptitude Index	−0423	−0679	−0274	−0627	−0600	−0375
General Aptitude Index	−0220	−0964	−0175	−0598	−0549	−0346
Electronic Aptitude Index	−0288	−0815	−0156	−0267	−0281	−0046
Coded 1 if male/ 0 if female or unknown	–	–	–	–	–	–
Coded 1 if female/ 0 if male or unknown	–	–	–	–	–	–
Age at time of enlistment	0681	0494	0743	0399	0653	0213
Education level at time of enlistment	0237	0026	0352	0240	0001	0088

[a]Decimal points have been omitted.

Table 7. **Correlations with Performance Ratings AFSC 431X1C**

N = 487

Predictor Variable	Criterion[a] General 1	Quantity 2	Quality 3	Exceeds Share 4	Self Motivating 5	Shares Knowledge 6
Grade	2137	1324	1900	1434	1790	1606
Total active federal military service, TAFMS	1338	0798	1290	0818	1053	1324
Decoding test score	−0069	−0313	−0460	−0086	−0443	−0181
Memory for landmarks test score	0898	0778	0492	0957	0840	0751
Complex scale reading test score	−0111	−0187	0121	−0069	−0137	−0144
Pursuit test score	−0288	−0438	−0190	−0419	−0208	−0275
Figure analogies test score	−0348	−0455	−0513	−0313	−0257	−0148
Hands test score	0745	0708	0757	0412	0698	0699
Cubes test score	−0001	−0108	−0254	−0047	−0191	−0074
Mechanical principles test score	0438	−0065	0315	0417	0286	0369
Following directions test score	0502	0329	0499	0378	0420	0286
Practical estimations test score	0734	0414	0589	0551	0483	0576
Spatial reasoning test score	0414	0368	0613	0459	0332	0029
Coded 1 if married/ 0 if otherwise or unknown	1026	0675	0469	0314	0564	0622
Coded 1 if divorced or separated/ 0 if otherwise, unk.	−0695	−0608	−0656	−0344	−0553	−0201
Code 1 if single/ 0 if otherwise or unknown	−0808	−0445	−0273	−0130	−0363	−0545
Size of city of origin	0080	−0185	−0289	0318	0447	0294
Mechanical Aptitude Index	0612	0900	0706	0796	0669	0498
Administrative Aptitude Index	0184	0244	0349	0362	0293	0298
General Aptitude Index	0232	0599	0319	0568	0510	0241
Electronic Aptitude Index	0632	0645	0918	0859	0878	0768
Coded 1 if male/ 0 if female or unknown	–	–	–	–	–	–
Coded 1 if female/ 0 if male or unknown	–	–	–	–	–	–
Age at time of enlistment	1245	0727	1378	1102	1293	1532
Education level at time of enlistment	0589	0594	0637	0335	0444	0579

[a]Decimal points have been omitted.

Comparison R^2s are presented in Table 8 for blocks of data predicting each of the six criteria for each AFSC. The task rating row entries of Table 8 can be compared directly to the R^2s of Table 4. Large

Table 8. **Regression Problems Compared for Three AFSCs – R^2s**

Criterion	Predictor System[a]	Number of Predictors[b]			AFSC — R^2		
		291X0	304X4	431X1C	291X0	304X4	431X1C
General performance rating	All variables	127	215	135	.3402	.3837	.3287
	Task ratings	102	190	110	.2814	.3319	.2707
	Demographic	10	10	10	.0287	.0227	.0568
	Aptitude	15	15	15	.0463	.0426	.0272
Amount of work performed rating	All variables	127	215	135	.3394	.4449	.3280
	Task ratings	102	190	110	.2915	.3721	.2995
	Demographic	10	10	10	.0308	.0133	.0250
	Aptitude	15	15	15	.0368	.0261	.0321
Quality of work rating	All variables	127	215	135	.3397	.4126	.3050
	Task ratings	102	190	110	.2840	.3483	.2403
	Demographic	10	10	10	.0304	.0187	.0463
	Aptitude	15	15	15	.0424	.0306	.0413
Will work more than required rating	All variables	127	215	135	.3494	.4520	.3241
	Task ratings	102	190	110	.3024	.4038	.2660
	Demographic	10	10	10	.0424	.0087	.0293
	Aptitude	15	15	15	.0288	.0421	.0275
Self-initiating (needs no prodding) rating	All variables	127	215	135	.3741	.3892	.3258
	Task ratings	102	190	110	.3304	.3384	.2686
	Demographic	10	10	10	.0346	.0128	.0441
	Aptitude	15	15	15	.0269	.0315	.0287
Will share knowledge and skill rating	All variables	127	215	135	.3666	.3058	.3015
	Task ratings	102	190	110	.3317	.2821	.2647
	Demographic	10	10	10	.0343	.0210	.0359
	Aptitude	15	15	15	.0184	.0352	.0223

Note. — Number of observations per coefficient = number of incumbent ratees = AFSC 291X0, 457; AFSC 304X4, 399; AFSC 431X1C, 487.

[a]Inclusion of a membership dummy variable (not rated on task) doubles the number of predictors.

[b]Designating 10 demographic variables is arbitrary: the male/female coding results in one nonredundant variable for AFSC 291X0 and no new variable for AFSCs 304X4 and 431X1C. Problems were computed before it was learned that full data existed for demographic variables and that the unknown coding was empty.

reductions in sample Ns have taken place due to dropping incumbents without test data; AFSC 291X0 is down from 1,254 to 457; AFSC 304X4 down from 806 to 399; and AFSC 431X1C down from 1,195 to 487. Increases in R^2 have occurred, roughly of values .07 or .08. These are appreciable amounts. Their size should serve to caution one against liberally interpreting R^2 values when large numbers of predictors are used with relatively small Ns. However, comparison of the values in Table 8 with those in Table 4 reveals that there have been only minor changes in the patterns of the three AFSCs. Prediction of overall ratings for AFSC 304X4 remains the best, and AFSC 291X0 predictions are slightly better than those for AFSC 431X1C. The differences between the latter two are small. The relative order of prediction among the six criteria within each AFSC is not entirely consistent between the two tables. However, all these differences are trivial compared to the size of the prediction difference between task prediction and that available from demographic and test data. The R^2s from these data of record are about one-tenth the values of the task prediction R^2s.

With few exceptions the R^2s of Table 8 drop in direct correspondence to the number of predictors in the system. It might be concluded that predictiveness of task ratings, taken item-by-item is not greater than that of the data of record. This is true for a single task when zeros have been introduced into the rating vector to include persons who were not rated on the performance of that task. However, by consulting Figures 4 and 5 it will be seen that when task ratings were predicting only overall performance of incumbents who performed those tasks, the correlations ranged from .40 to .60. Squared, any one of these vectors could account for from 16 to 30 percent of the criterion variance. Table 8 consequently presents an ambiguous situation in which the higher R^2s accompanying task ratings could be attributable either to numerous predictors or to substantial relationships between task performance and overall performance ratings. It should be pointed out that while all correlations between task performance ratings and overall ratings are positive in Figures 4 and 5, in the matrices involving incumbents who did not perform tasks there are many significant negative correlations between the task rating vectors and the criterion vectors. Thus, the nature of the buildup of R^2 is a very complex interative process in which details cannot be determined by mere comparison of R^2 values.

Each kind of predictor, task performance ratings, demographic data, and aptitude measures was combined in a single equation to predict each of the six criterion dimensions per specialty in Table 8. Then each block was removed to leave the remaining two blocks of predictors in Tables 9, 10, and 11. All full and restricted models for an AFSC appear in one table. F-tests were applied to determine the statistical

Table 9. **Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 291X0**

Number of observations = 457; number of task predictors = 51 tasks + 51 dummy membership variables = 102; number of demographic variables = 10; number of aptitude measures = 15.

Degrees of freedom,[a] by problem: task removed, $Df_1 = 102$, $Df_2 = 330$; demographic removed, $Df_1 = 10$, $Df_2 = 330$; aptitude measures removed, $Df_1 = 15$, $Df_2 = 330$.[b]

Criterion	Problem	R^2 Full	R^2 Restricted	Difference	F-ratio	Probability
General performance rating	Remove task variables	.3402	.0723	.2679	1.3133	.039
	Remove demographics	.3402	.3194	.0207	1.0[illegible]	.412
	Remove aptitude variables	.3402	.3033	.0368	1.228[illegible]	.248
Amount of work rating	Remove task variables	.3394	.0656	.2738	1.3411	.029
	Remove demographics	.3394	.3160	.0234	1.1701	.310
	Remove aptitude variables	.3394	.3092	.0302	1.0075	.447
Quality of work rating	Remove task variables	.3397	.0727	.2670	1.3085	.041
	Remove demographics	.3397	.3153	.0245	1.2224	.275
	Remove aptitude variables	.3397	.3066	.0331	1.1031	.352
Exceeds share rating	Remove task variables	.3494	.0734	.2760	1.3724	.020
	Remove demographics	.3494	.3235	.0259	1.3158	.220
	Remove aptitude variables	.3494	.3311	.0184	.6206	.858
Self-initiating rating	Remove task variables	.3741	.0606	.3135	1.6207	.001
	Remove demographics	.3741	.3586	.0155	.8167	.613
	Remove aptitude variables	.3741	.3701	.0040	.1412	1.000
Shares knowledge rating	Remove task variables	.3666	.0538	.3128	1.5978	.001
	Remove demographics	.3666	.3494	.0172	.8971	.536
	Remove aptitude variables	.3666	.3454	.0212	.7358	.748

[a]An arbitrary decision was made in the case of the demographic variables to use 10 as the base; since there is redundancy in male/female and married/separated/single, the number of predictor variables has been overestimated, and upon correction might show significance. Earlier problems permitting the number of variables actually entering the equation as base were discarded because they were suspected of overestimating a significances. Recomputation can be made with 9 demographic variables for AFSC 291X0, and 8 for the other two AFSCs because there were no women.

[b]See footnote of 1 of Table 11; correct to 51 tasks.

Table 10. **Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 304X4[a]**

Number of observations = 399; number of task rating predictors = 95 + 95 dummy membership variables = 190; number of demographic predictors = 10; number of aptitude predictors = 15.

Degrees of freedom by problem:[b] tasks removed, $Df_1 = 190$, $Df_2 = 189$; demographics removed, $Df_1 = 10$, $Df_2 = 189$; aptitude variables removed, $Df_1 = 15$, $Df_2 = 189$.

Criterion	Problem	R^2 Full	R^2 Restricted	Difference	F-ratio	Probability
General performance rating	Remove task variables	.3837	.0543	.3294	.5490	1.000
	Remove demographics	.3837	.3659	.0178	.5465	.855
	Remove aptitude variables	.3837	.3486	.0351	.7177	.765
Amount of work rating	Remove task variables	.4449	.0393	.4056	.7506	.974
	Remove demographics	.4449	.4294	.0155	.5271	.870
	Remove aptitude variables	.4449	.4233	.0216	.4913	.943
Quality of work rating	Remove task variables	.4126	.0446	.3680	.6436	.999
	Remove demographics	.4126	.3579	.0547	1.7612	.070
	Remove aptitude variables	.4126	.3889	.0237	.5094	.934
Exceeds share rating	Remove task variables	.4520	.0489	.4031	.7555	.972
	Remove demographics	.4520	.4361	.0159	.5484	.854
	Remove aptitude variables	.4520	.3956	.0563	1.2954	.208
Self-initiating rating	Remove task variables	.3892	.0390	.3503	.5891	1.000
	Remove demographics	.3892	.3779	.0114	.3521	.965
	Remove aptitude variables	.3892	.3795	.0097	.2010	1.000
Shares knowledge rating	Remove task variables	.3058	.0503	.2555	.3780	1.000
	Remove demographics	.3058	.2891	.0167	.4545	.917
	Remove aptitude variables	.3058	.2847	.0211	.3826	.982

[a]See footnote 1 of Table 11; correct to 95 tasks.

[b]See footnote 1 of Table 9.

Table 11. **Unique Contributions of Blocks of Variables to Predicting Overall Performance Ratings, AFSC 431X1C[a]**

Number of observations = 487; number of task rating predictors = 55 + 55 dummy membership variables = 110; number of demographic predictors = 10; number of aptitude predictors = 15.

Degrees of freedom by problem:[b] tasks removed, $Df_1 = 110$, $Df_2 = 352$; demographics removed, $Df_1 = 10$, $Df_2 = 352$; aptitude variables removed, $Df_1 = 15$, $Df_2 = 352$.

Criterion	Problem	R^2 Full	R^2 Restricted	Difference	F-ratio	Probability
General performance rating	Remove task variables	.3287	.0794	.2493	1.1885	.123
	Remove demographics	.3287	.2879	.0408	2.1391	.021
	Remove atptitude variables	.3287	.3061	.0226	.7898	.689
Amount of work rating	Remove task variables	.3280	.0514	.2766	1.3169	.032
	Remove demographics	.3280	.3152	.0128	.6689	.753
	Remove aptitude variables	.3280	.3115	.0164	.5743	.894
Quality of work rating	Remove task variables	.3050	.0768	.2281	1.0503	.365
	Remove demographics	.3050	.2850	.0200	1.0132	.432
	Remove aptitude variables	.3050	.2768	.0281	.9498	.509
Exceeds share rating	Remove task variables	.3241	.0534	.2707	1.2814	.048
	Remove demographics	.3241	.2820	.0421	2.1905	.018
	Remove aptitude variables	.3241	.2852	.0389	1.3490	.171
Self-initiating rating	Remove task variables	.3258	.0669	.2589	1.2288	.083
	Remove demographics	.3258	.2935	.0323	1.6862	.082
	Remove aptitude variables	.3258	.2984	.0274	.9547	.503
Shares knowledge rating	Remove task variables	.3015	.0543	.2472	1.1327	.200
	Remove demographics	.3015	.2738	.0278	1.3994	.179
	Remove aptitude variables	3015	.2688	.0327	1.0990	.356

[a]The full model contains a block of 55 task ratings and 55 dummy variables, a block of 10 demographic variables, and a block of 15 aptitude measures; restricted models were obtained by removing one of the three blocks from the full model.

[b]See footnote 1 of Table 9.

significance of the loss due to removal of a block of predictors. In Tables 9, 10, and 11 the full model equation contains all predictors and the restricted model contains the remaining two blocks of predictors. This is the unique contribution of one kind of predictor made in the presence of the other two. The significance is in terms of probability, the number of times in 1,000 trials that the difference in R^2 between the full and restricted models was attributable to chance.

From Table 9, it appears that the contribution of task ratings to the prediction of AFSC 291X0 performance ratings was significant statistically for all six criteria. However, the number of predictors is large and the number of observations per correlation coefficient is relatively small, which is a condition that suggests that the results should be treated cautiously. Aptitude data did not make a contribution to prediction in the presence of both task and demographic variables, and demographic data made no contribution in the presence of both task and aptitude data.

From Table 10, it is plain that very large losses, as high as 40 percent of the criterion accountability, are nonsignificant. No difference value is significant in Table 10. The reason is the high ratio of criterion predictors (95 tasks, or 190 predictors) to the number of performance observations per correlation coefficient.

In Table 11, there are only two criteria which reflect significant unique prediction from task ratings for AFSC 43;X1C. The significance is marginal. However, in this AFSC the demographic data made significant contributions to predicting ratings on two criteria, general performance and willingness to do more than one's share. Aptitude measures made no significant contributions in the presence of both task and demographic variables. Again, very large reductions in R^2 can be statistically nonsignificant when the number of predictors removed from the system is high in relation to the number of observations per correlation coefficient. When tasks were removed, a loss of 25 percent of the variance prediction was nonsignificant, although removal of the block of demographic variables caused the loss of only 4 percent prediction, which was significant.

Reduction of Five Task Performance Predictors

All in all, the findings from inclusion of all task ratings are suggestive hypothesized relationships, but the results presented in Tables 9 through 11 are not clearly delineated. The obvious solution to the problem was to cut down the number of task rating predictors. This was done, and the results appear in Table 12 and following.

A sequential set of five tasks was chosen in advance of computing regression problems by dividing the number of tasks into five equal intervals, beginning with the second task rated. This concentrated tasks early in the list, which was intentional, because it was suspected that raters may have given their best efforts during the first period of the survey. This set of five tasks is here termed the sequential set; a second set of five tasks was chosen on the basis of their probable predictiveness, and these are termed the critical set.

Five tasks were selected in order to make the number of predictors the same as that of the smallest block of nontask data, the set of 10 demographic variables. This created a new problem because few, if any, tasks in AFSC 291X0 were performed by as many as one-half of the incumbents. In AFSC 304X4 there were some tasks on which less than 10 percent of the incumbents were rated. Unless an effort was made to find five popularly rated tasks, it would be possible for an incumbent to enter the regression problems by being a nonperformer on every task. Accordingly, for the problems using carefully selected tasks, two bases of selection were applied jointly, the percent of incumbents rated on the task and the point of entry of the task rating variable into the iterations of a regression problem. The problem used for that purpose was the prediction of general performance from all variables, appearing in the first line of Table 8. In general, the membership variable entered the iterations in close proximity to the task rating variable, but not always.

Zero-order correlations are given for both sets of five tasks and membership variables against the criterion of general overall performance in Table 12. The membership variable is labeled "Dummy." The intercorrelations are not presented in Table 12; they were quite substantial, ranging from .44 to .89, or from −.44 to −.89, for different tasks and dummies. The correlation between a task rating vector and its

Table 12. **Statistics of Selected Tasks**

5 Predictive Tasks				5 Sequence-Selected Tasks			
List Number	Correlation with General Performance: Task	Correlation with General Performance: Dummy	% Rated on Task	List Number	Correlation with General Performance: Task	Correlation with General Performance: Dummy	% Rated on Task
			AFSC 291X0				
02	.0801	−.0128	31	02	.0801	−.0128	31
13	−.0244	.0685	39	12	−.0152	.0763	36
24	.0089	.0297	26	22	.0040	.0387	34
40	−.0695	.1215	24	32	.0209	.0108	24
49	−.0030	.0470	36	42	−.0387	.1045	24
			AFSC 304X4				
04	.1274	−.0722	29	02	.1293	−.0569	37
21	−.0303	.0709	50	20	.0548	.0138	46
67	.0926	−.0236	50	38	.0173	.0315	
72	.0628	.0152	50	56	.0417	.0200	46
88	.0098	.0504	44	68	0796	−.0134	49
			AFSC 431X1C				
02	.0299	.0573	51	02	.0299	.0573	51
04	.1433	−.0906	33	13	.0580	.0049	43
11	.0196	.0530	52	24	.1060	−.0845	27
31	.0356	.0207	48	35	.0469	.0070	46
40	.0046	.0455	46	46	.0157	.0427	43

dummy lay between −.94 and −.995. It is likely that clusters of tasks corresponded to certain jobs, which helped to create a correlation that depended upon the fact of being rated on a task rather than upon the performance rating level. This should be borne in mind when interpreting these analyses.

The sequential sets of tasks began with the second task. When the critical sets were chosen, it turned out that the second task was one of the best predictors, which resulted in its appearing in both the sequential list and the critical list for two of the AFSCs. Otherwise, the sets are not duplicative. Neither set contains any of the very rare tasks, and it seems probable that most of the incumbents were represented by a task performance rating at least once in each set of computations. This, however, is a presumption, and it may not hold true for AFSC 304X4, which had 95 tasks to be spread over 399 ratees.

Tables 13 through 17 afford comparisons of the three AFSCs. The blocks of task performance predictors contain 10 predictors, while the 10 demographic and 15 aptitude variables are unchanged from the previous analyses. The results of the analyses shown in Table 13 are conclusive. The evidence that task performance ratings predicted overall performance variance in a way that aptitude and/or demographic data could not is inescapable. In all problems in the table the likelihood that any difference between the full and restricted models could have occurred by chance is less than 1 in 1,000 trials.

Table 14 gives the unique contribution of demographic data to each of the six criteria, and Table 15 does the same for aptitude data. In the presence of task ratings combined with the other predictor block neither demographic nor aptitude measures contribute much toward predicting overall performance variance, generally. Nevertheless, there are suggestions of unique contribution to motivational performance ratings for AFSC 291X0 from demographic data, and correspondingly, contributions to general performance ratings for AFSC 431X1C. No statistically significant contributions were made by the block of aptitude variables reported in Table 15.

Tables 16 and 17 were developed to test the possibility that overlap between demographic and aptitude data was obscuring the contribution of some variables of record toward predicting overall

Table 13. **Unique Prediction Contribution of 5 Task Ratings to Overall Performance[a, b]**

Criterion	AFSC 291X0 $Df_1 = 10$ $Df_2 = 422$					AFSC 304X4 $Df_1 = 10$ $Df_2 = 364$					AFSC 431X1C $Df_1 = 10$ $Df_2 = 452$				
	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.
5 Tasks Selected for Predictiveness and Percent Performing															
General performance rating	.2523	.0723	.1800	10.1582	.000	.2313	.0543	.1770	8.3799	.000	.2564	.0794	.1770	10.7579	.000
Amount of work rating	.2468	.0656	.1812	10.1502	.000	.1968	.0393	.1575	7.1391	.000	.2435	.0514	.1921	11.4772	.000
Quality of work rating	.2492	.0727	.1765	9.9198	.000	.2164	.0446	.1718	7.9820	.000	.2440	.0768	.1671	9.9925	.000
Exceeds share rating	.2547	.0734	.1812	10.2605	.000	.2068	.0489	.1579	7.2465	.000	.2320	.0534	.1786	10.5104	.000
Self-initiating rating	.2759	.0606	.2153	12.5479	.000	.1971	.0390	.1582	7.1712	.000	.2298	.0669	.1628	9.5559	.000
Shares knowledge rating	.2311	.0538	.1772	9.7265	.000	.1824	.0503	.1320	5.8785	.000	.2055	.0543	.1512	8.6024	.000
5 Tasks Selected By Sequence															
General performance rating	.2221	.0723	.1498	8.1268	.000	.1980	.0543	.1438	6.5246	.000	.2157	.0794	.1363	7.8573	.000
Amount of work rating	.1989	.0656	.1332	7.0184	.000	.1917	.0393	.1524	6.8632	.000	.1888	.0514	.1374	7.6563	.000
Quality of work rating	.2257	.0727	.1530	8.3367	.000	.2116	.0446	.1670	7.7084	.000	.2145	.0768	.1376	7.9197	.000
Exceeds share rating	.2072	.0734	.1337	7.1170	.000	.2032	.0489	.1543	7.0511	.000	.1897	.0534	.1363	7.6013	.000
Self-initiating rating	.2272	.0606	.1666	9.0981	.000	.2024	.0390	.1634	7.4586	.000	.1815	.0669	.1146	6.3283	.000
Shares knowledge rating	.2062	.0538	.1524	8.1027	.000	.1760	.0503	.1257	5.5512	.000	.1711	.0543	.1168	6.3664	.000

[a]The full model predictor system contains 5 task ratings and 5 dummy variables, 10 demographic variables, and 15 aptitude variables; the restricted model contains 10 demographic variables and 15 aptitude variables.

[b]See footnote 1 of Table 9.

Table 14. **Unique Demographic Contribution to Aptitude and 5 Task Overall Performance Prediction[a, b]**

Criterion	AFSC 291X0 $Df_1 = 10$ $Df_2 = 422$					AFSC 304X4 $Df_1 = 10$ $Df_2 = 364$					AFSC 431X1C $Df_1 = 10$ $Df_2 = 452$				
	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.
5 Tasks Selected for Predictiveness and Percent Performing															
General performance rating	.2523	.2286	.0237	1.3367	.208	.2313	.2025	.0288	1.3626	.196	.2564	.2237	.0327	1.9868	.033
Amount of work rating	.2468	.2238	.0230	1.2906	.233	.1968	.1749	.0219	.9936	.448	.2435	.2352	.0082	.4927	.895
Quality of work rating	.2492	.2360	.0132	.7429	.684	.2164	.1975	.0189	.8785	.553	.2440	.2141	.0299	1.7879	.060
Exceeds share rating	.2547	.2167	.0379	2.1487	.020	.2068	.1903	.0165	.7570	.670	.2320	.2158	.0162	.9549	.482
Self-initiating rating	.2759	.2331	.0428	2.4943	.006	.1971	.1801	.0170	.7707	.657	.2298	.2021	.0277	1.6268	.096
Shares knowledge rating	.2311	.2037	.0274	1.5019	.136	.1824	.1726	.0098	.4347	.929	.2055	.1795	.0260	1.4805	.144
5 Tasks Selected By Sequence															
General performance rating	.2221	.2018	.0203	1.1007	.360	.1980	.1796	.0185	.8387	.591	.2157	.1862	.0295	1.7007	.078
Amount of work rating	.1989	.1881	.0107	.5653	.842	.1917	.1775	.0142	.6411	.778	.1888	.1768	.0120	.6688	.754
Quality of work rating	.2257	.2042	.0215	1.1699	.309	.2116	.2031	.0084	.3898	.951	.2145	.1900	.0245	1.4094	.173
Exceeds share rating	.2072	.1741	.0330	1.7575	.066	.2032	.1970	.0062	.2814	.985	.1897	.1729	.0168	.9344	.501
Self-initiating rating	.2272	.1945	.0326	1.7823	.062	.2024	.1956	.0068	.3121	.978	.1815	.1562	.0254	1.4010	.177
Shares knowledge rating	.2062	.1887	.0175	.9308	.505	.1760	.1658	.0101	.4483	.922	.1711	.1429	.0282	1.5361	.124

[a]The full model predictor system contains 5 task ratings and 5 dummy variables, 10 demographic variables, and 15 aptitude variables; the restricted model contains 10 task variables and 15 aptitude variables.

[b]See footnote 1 of Table 9.

Table 15. **Unique Aptitude Contribution to Demographic and 5 Task Overall Performance Prediction[a, b]**

Criterion	AFSC 291X0 $Df_1 = 15$ $Df_2 = 422$					AFSC 304X4 $Df_1 = 15$ $Df_2 = 314$					AFSC 431X1C $Df_1 = 15$ $Df_2 = 452$				
	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.
5 Tasks Selected for Predictiveness and Percent Performing															
General performance rating	.2523	.2153	.0370	1.3912	.147	.2313	.1909	.0404	1.2742	.216	.2564	.2361	.0202	.8194	.656
Amount of work rating	.2468	.2200	.0268	1.0029	.451	.1968	.1664	.0305	.9206	.541	.2435	.2185	.0250	.9947	.459
Quality of work rating	.2492	.2231	.0261	.9779	.478	.2164	.1910	.0255	.7888	.690	.2440	.2120	.0320	1.2764	.213
Exceeds share rating	.2547	.2350	.0197	.7434	.740	.2068	.1688	.0380	1.1624	.299	.2320	.2093	.0227	.8894	.576
Self-initiating rating	.2789	.2600	.0159	.6178	.861	.1971	.1693	.0278	.8400	.633	.2298	.2025	.0272	1.0655	.387
Shares knowledge rating	.2311	.2193	.0118	.4307	.970	.1824	.1548	.0276	.8177	.658	.2055	.1806	.0249	.9461	.512
5 Tasks Selected By Sequence															
General performance rating	.2221	.1843	.0378	1.3676	.159	.1980	.1605	.0375	1.1358	.322	.2157	.2005	.0152	.5824	.889
Amount of work rating	.1989	.1797	.0192	.6748	.810	.1917	.1639	.0278	.8347	.639	.1888	.1645	.0243	.9041	.560
Quality of work rating	.2257	.1872	.0385	1.3990	.144	.2116	.1824	.0292	.8982	.566	.2145	.1868	.0277	1.0635	.389
Exceeds share rating	.2072	.1838	.0233	.8282	.646	.2032	.1696	.0336	1.0236	.430	.1897	.1691	.0206	.7662	.716
Self-initiating rating	.2272	.2094	.0178	.6478	.835	.2024	.1798	.0226	.6867	.798	.1815	.1561	.0254	.9347	.525
Shares knowledge rating	.2062	.1976	.0086	.3054	.995	.1760	.1497	.0263	.7735	.707	.1711	.1536	.0174	.6334	.848

[a]The full model predictor system contains 5 task ratings and 5 dummy variables, 10 demographic variables, and 15 aptitude variables; the restricted model contains 10 task variables and 10 demographic variables.

[b]See footnote 1 of Table 9.

Table 16. **Demographic Contribution to Performance Prediction by 5 Tasks[a, b]**

Criterion	AFSC 291X0 $Df_1 = 10$ $Df_2 = 437$					AFSC 304X4 $Df_1 = 10$ $Df_2 = 379$					AFSC 431X1C $Df_1 = 10$ $Df_2 = 467$				
	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.
5 Tasks Selected for Predictiveness and Percent Performing															
General performance rating	.2153	.1905	.0249	1.3843	.185	.1909	.1566	.0343	1.6082	.102	.2361	.2008	.0354	2.1613	.019
Amount of work rating	.2200	.1926	.0274	1.5340	.124	.1664	.1465	.0199	.9051	.528	.2185	.2067	.0118	.7070	.718
Quality of work rating	.2231	.1891	.0340	1.9130	.042	.1910	.1698	.0211	.9894	.452	.2120	.1801	.0319	1.8908	.044
Exceeds share rating	.2350	.1957	.0393	2.2445	.015	.1688	.1516	.0171	.7809	.647	.2093	.1924	.0169	.9998	.442
Self-initiating rating	.2600	.2229	.0371	2.1915	.017	.1693	.1514	.0180	.8203	.609	.2025	.1776	.0249	1.4578	.152
Shares knowledge rating	.2193	.1895	.0298	1.6661	.086	.1548	.1362	.0186	.8342	.596	.1806	.1546	.0259	1.4779	.144
5 Tasks Selected By Sequence															
General performance rating	.1843	.1632	.0211	1.1322	.336	.1605	.1431	.0174	.7842	.644	.2005	.1658	.0347	2.0290	.029
Amount of work rating	.1797	.1589	.0208	1.1077	.355	.1639	.1510	.0129	.5845	.827	.1645	.1493	.0152	.8498	.581
Quality of work rating	.1872	.1609	.0263	1.4114	.172	.1824	.1718	.0106	.4901	.896	.1868	.1561	.0306	1.7598	.066
Exceeds share rating	.1838	.1514	.0325	1.7381	.070	.1696	.1601	.0095	.4329	.930	.1691	.1639	.0051	.2890	.984
Self-initiating rating	.2094	.1738	.0356	1.9650	.036	.1798	.1711	.0087	.4025	.945	.1561	.1313	.0249	1.3768	.188
Shares knowledge rating	.1976	.1736	.0240	1.3070	.224	.1497	.1329	.0168	.7503	.677	.1536	.1231	.0305	1.6854	.081

[a]The full model predictor system contains 5 task ratings and 5 dummy variables; plus 10 demographic measures, the restricted model contains the 5 task and dummy variables.

[b]See footnote 1 of Table 9.

Table 17. **Aptitude Contribution to Performance Prediction by 5 Tasks[a]**

Criterion	AFSC 291X0 $Df_1 = 15$ $Df_2 = 432$					AFSC 304X4 $Df_1 = 15$ $Df_2 = 374$					AFSC 431X1C $Df_1 = 15$ $Df_2 = 462$				
	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.	R^2 Full	R^2 Rest	Diff	F-rat.	Prob.
5 Tasks Selected for Predictiveness and Percent Performing															
General performance rating	.2286	.1905	.0381	1.4243	.132	.2025	.1566	.0459	1.4357	.128	.2237	.2008	.0229	.9080	.555
Amount of work rating	.2238	.1926	.0312	1.1575	.303	.1749	.1465	.0285	.8599	.610	.2352	.2067	.0286	1.1501	.309
Quality of work rating	.2360	.1891	.0469	1.7675	.037	.1975	.1698	.0277	.8600	.610	.2141	.1801	.0340	1.3334	.178
Exceeds share rating	.2167	.1957	.0210	.7735	.707	.1903	.1516	.0386	1.1894	.277	.2158	.1924	.0234	.9179	.544
Self-initiating rating	.2331	.2229	.0102	.3835	.983	.1801	.1514	.0288	.8750	.593	.2027	.1776	.0244	.9421	.517
Shares knowledge rating	.2037	.1895	.0142	.5125	.934	.1726	.1362	.0364	1.0966	.358	.1795	.1546	.0249	.9329	.527
5 Tasks Selected By Sequence															
General performance rating	.2018	.1632	.0387	1.3949	.145	.1796	.1431	.0364	1.1070	.348	.1862	.1658	.0204	.7714	.710
Amount of work rating	.1881	.1589	.0293	1.0385	.414	.1775	.1510	.0265	.8021	.676	.1768	.1493	.0275	1.0304	.422
Quality of work rating	.2042	.1609	.0433	1.5668	.079	.2031	.1718	.0313	.9798	.476	.1900	.1561	.0339	1.2881	.205
Exceeds share rating	.1741	.1514	.0228	.7945	.684	.1970	.1601	.0369	1.1469	.312	.1729	.1639	.0090	.3349	.992
Self-initiating rating	.1945	.1738	.0207	.7403	.743	.1956	.1711	.0244	.7576	.725	.1562	.1313	.0249	.9087	.558
Shares knowledge rating	.1887	.1736	.0151	.5363	.920	.1658	.1329	.0329	.9848	.470	.1429	.1231	.0198	.7114	.774

[a]The full model predictor system contains 5 task ratings and 5 dummy variables, plus 15 aptitude measures; the restricted model contains the 5 task and dummy variables

performance. In these tables the other set of variables is screened out. Table 16 reaffirms the findings of Table 14 that demographic data make a contribution to predicting motivational rating variance for AFSC 291X0, and to predicting general performance ratings for AFSC 431X1C. There is a marginally significant indication that quality of work ratings in AFSC 431X1C were predicted by demographic data. The tasks selected by sequence reflect the same patterns as those selected for their probable predictiveness, but with slightly less statistical significance. In Table 17, there is only one statistically significant aptitude prediction, and it is marginal. However, failure of a block of aptitude predictors to make a significant contribution does not necessarily mean that an individual experimental test would also fail when used alone with task ratings as a baseline.

Cross-Rater Prediction of Overall Performance

Could rater tendency have been responsible for the obtained significances of contribution of task ratings toward predicting overall performance? The same individuals rated an incumbent's task and overall performance, and averaging two raters may not have eliminated the rater tendency effect. The solution was to compute cross-rater regression problems, which meant abandoning the supplementary sample, which had only one rater. The resulting Ns were drastically cut, eventuating in 170 cases for AFSC 291X0, 168 for AFSC 304X4, and 193 for AFSC 431X1C. All incumbents had task ratings from both raters, but not necessarily tasks in the two lists selected for regression problems. In fact, breaking up mean scores may have created a number of new cases with no rated tasks. The cross-rater computations shown in Table 18 employ the task performance ratings made by Rater A to predict the overall performance ratings made by Rater B, and vice versa. The significance of loss on removal of the task block follows the format of Table 13. The loss significance findings negate the hypothesis that rater tendency was responsible for the contribution of the task rating variables toward predicting overall performance ratings. If one assumes that rater tendencies are uncorrelated between raters, pronounced rater tendencies would lower cross-correlations, and the correlations might be lower than those from which Table 13 was computed. In fact, the correlations producing Table 18 were about the same, and the R^2s of Table 18 are considerably higher than those shown in Table 13. This is probably attributable to the relative size of the number of predictor variables to the reduced size of the number of observations per correlation coefficient. Table 18 is presented as a test of the rater tendency hypothesis, and it should not be quoted as a reflection of the entire study, which contains an additional sample.

Is it possible that the predictions found here are attributable to the incumbents who were not rated on any of the five tasks in each equation? There are four sets to be considered in identifying such cases: Rater A and Rater B, each rating a set of five critical and a set of five sequential tasks. A count of the cases with zero rating data was made, and the median number of each of the four conditions was found for each AFSC. Converted to percentages, the medians were approximately 12 percent for AFSC 291X0, 5 percent for AFSC 304X4, and 1 percent for AFSC 431X1C. It is highly unlikely that incumbents who lacked task ratings altogether could have greatly influenced the results for AFSCs 304X4 and 431X1C, although they may have had some effect in AFSC 291X0. The influence of the large group of incumbents who were nonrated on any specific task is a different matter; they could greatly influence the R^2. If nonrated incumbents had higher overall performance scores than their accompanying rated incumbents, the result would be negative r's, such as the one shown for task 40 under AFSC 291X0 in Table 12. Here the task rating predictor correlated −.07 with the criterion and the dummy correlated +.12.

In sum, the tests demonstrated that the large contribution of task performance ratings toward predicting overall performance was not attributable to rater tendency. Further, a simple count showed that it is highly improbable that nonperformers on every task were the source of the contribution of task ratings to R^2. One attractive explanation for the findings is that they were produced by a combination of rated performance quality on difficult tasks and ratings at the top of the scale on jobs composed of simple tasks.

Contributions of Single Variables of Data of Record

Table 19 concludes the analyses by returning to the data on which Tables 13 through 17 are based. In Table 19, the individual contribution of each of the 25 predictor variables representing data which were

Table 18. **Cross-Rater Regression Contributions from Sets of 5 Task Performance Ratings**

Criterion System[a]	AFSC 291X0 N = 170 df_1 = 10; df_2 = 135					AFSC 304X4 N = 148 df_1 = 10; df_2 = 113					AFSC 431X1C N = 193 df_1 = 10; df_2 = 158				
	Full Model R^2	Rest. Model R^2	Diff. R^2	F-Ratio	Prob.[b]	Full Model R^2	Rest. Model R^2	Diff. R^2	F-Ratio	Prob[b]	Full Model R^2	Rest. Model R^2	Diff. R^2	F-Ratio	Prob[b]
Tasks Selected for Criticality															
1. A on B	.2561	.1034	.1527	2.771	.004	.2933	.1246	.1687	2.698	.005	.2291	.1462	.0829	1.698	.085
B on A	.3599	.2189	.1409	2.972	.002[c]	.2632	.1773	.0859	1.317	.230[c]	.2189	.1278	.0912	1.844	.057
2. A on B	.2875	.1591	.1284	2.433	.011	.2959	.1389	.1570	2.519	.009	.2076	.1009	.1067	2.127	.025
B on A	.2741	.1807	.0934	1.736	.079[c]	.2268	.1181	.1087	1.588	.119[c]	.2478	.1293	.1186	2.490	.008[c]
3. A on B	.2748	.1328	.1421	2.645	.006	.2075	.1118	.0958	1.3656	.205	.2102	.1239	.0863	1.726	.079
B on A	.2311	.1512	.0799	1.404	.185[c]	.3000	.2031	.0968	1.563	.127	.2545	.1467	.1078	2.2850	.016
4. A on B	.2832	.1363	.1469	2.767	.004	.3367	.1573	.1794	3.057	.002	.2515	.1852	.0664	1.401	.184
B on A	.3029	.1795	.1234	2.390	.012[c]	.1948	.1115	.0833	1.169	.319[c]	.2926	.1786	.1140	2.546	.007[c]
5. A on B	.2393	.1693	.0934	1.736	.079	.3224	.1403	.1821	3.036	.002	.2265	.1604	.0661	1.350	.209
B on A	.2791	.1820	.0971	1.818	.063	.2943	.1674	.1270	2.033	.036[c]	.2854	.1610	.1244	2.750	.004[c]
6. A on B	.2619	.1220	.1389	2.538	.008	.2319	.1342	.0978	1.438	.173	.2229	.1209	.1020	2.073	.030
B on A	.2291	.1795	.0497	0.8696	.563[c]	.2183	.1718	.0465	0.672	.746	.1573	.1087	.0486	0.911	.525
Sequence Selected Tasks															
1. A on B	.2657	.1034	.1623	2.984	.002	.2775	.1246	.1528	2.390	.013	.2086	.1462	.0624	1.246	.266
B on A	.3469	.2189	.1280	2.646	.006[c]	.2954	.1773	.1181	1.895	.053[c]	.2650	.1278	.1373	2.951	.002[c]
2. A on B	.3015	.1591	.1424	2.751	.004	.2994	.1389	.1605	2.589	.007	.1922	.1009	.0913	1.787	.067
B on A	.3675	.1807	.1867	3.988	.000[c]	.2152	.1181	.0951	1.365	.206[c]	.2509	.1293	.1216	2.565	.007[c]
3. A on B	.2787	.1328	.1460	2.732	.004	.2601	.1118	.1483	2.265	.019	.2704	.1239	.1465	3.172	.001
B on A	.3825	.1512	.2313	5.057	.000[c]	.3119	.2031	.1088	1.787	.071	.2616	.1467	.1148	2.457	.009[c]
4. A on B	.2574	.1363	.1211	2.202	.021	.3267	.1573	.1694	2.844	.003	.2783	.1852	.0931	2.039	.033
B on A	.3592	.1795	.1797	3.787	.000[c]	.1720	.1115	.0605	0.825	.605[c]	.3109	.1786	1.323	3.032	.002[c]
5. A on B	.2868	.1693	.1175	2.224	.020	.2872	.1403	.1469	2.329	.016	.2724	.1604	.1120	2.432	.010
B on A	.2777	.1820	.0957	1.789	.068	.2509	.1674	.0835	1.260	.261	.3021	.1610	.1415	3.204	.001[c]
6. A on B	.2630	.1220	.1410	2.583	.007	.3195	.1342	.1853	3.077	.002	.2130	.1209	.0921	1.849	.056
B on A	.2992	.1795	.1197	2.306	.016[c]	.2312	.1718	.0594	0.873	.560[c]	.1742	.1087	.0654	1.252	.263

[a]Numbers 1–6 are the 6 rating dimensions shown in Table 1 and elsewhere. A on B means task performance ratings by raters A predicting overall ratings by raters B; B on A is the reverse. The full model contains 5 task rating predictors, 5 dummy predictors, and 25 variables of record; the last are retained.

[b]If the probability value for either member of a pair is not greater than .014 the cross-rater prediction can be considered to be significant. Pairings failing significance are: 2 for AFSC 291X0; 3 for AFSC 304X4; 4 for AFSC 431X1C.

[c]Considered significant as a pair.

Table 19. **Comparison of Single Variable Contributions to Overall Prediction by Task Ratings**

Predictor Variable	Task Set	AFSC 291X0[a]						AFSC 304X4[b]						AFSC 431X1C[c]					
		Gen'l	Quant.	Qual.	Exc'd Share	Self Motiv.	Shares Know.	Gen'l	Quant.	Qual.	Exc'd Share	Self Motiv.	Shares Know.	Gen'l	Quant.	Qual.	Exc'd Share	Self Motiv.	Shares Know.
Grade	Critical		.015					.006						<.001	.014	<.001	.010	.001	.002
	Sequenced		.025											<.001	.009	<.001		.001	<.001
Total active federal military service	Critical							.018						.002		.002		.017	.002
	Sequenced													.006		.007		.029	.001
Memory for landmarks test score	Critical							.023											
	Sequenced			.015															
Complex scale reading test score	Critical							.013		.031			.023						
	Sequenced							.008		.010	.021		.009						.027
Hands test score	Critical																		
	Sequenced														.002				
Spatial reasoning test score	Critical	.021		.024															
	Sequenced	.021		.024															
Coded 1 if married/ 0 other or unknown	Critical																		
	Sequenced				.046	.043													
Coded 1 if divorced or separated/ 0 other	Critical	.034	.047					.001	.012	.008	.015	.007							
	Sequenced	.011																	
Coded 1 if single/ 0 otherwise or unknown	Critical					.041													
	Sequenced																		
Mechanical Aptitude Index	Critical		.022			.030			.019							.048			
	Sequenced		.041						.040							.048			.019
General Aptitude Index	Critical								.008			.031							
	Sequenced								.008										
Coded 1 if male/0 if female or unknown	Critical				.009														
	Sequenced			.016	.008	.044													
Coded 1 if female/ 0 if male or unknown	Critical				.009														
	Sequenced			.016	.008	.044													
Age at time of enlistment	Critical													.004		.002	.017	.003	<.001
	Sequenced													.022		.007		.008	<.001
Education at time of enlistment	Critical																		
	Sequenced		.047																

[a] $df_1 = 1, df_2 = 446$.
[b] $df_1 = 1, df_2 = 388$.
[c] $df_1 = 1, df_2 = 476$.

either on record, or which might be obtained by testing the incumbent, is examined for prediction of the six overall criteria. All three AFSCs are presented in Table 19 with the criterion variables composing the columns. The values shown are probability estimates, the chances in 1,000 that the obtained difference between the full and restricted models could have arisen by chance. All values exceeding .049 have been dropped, leaving only prediction contributions that are significant at the 5 percent level of confidence or better. The full model contained five task performance ratings and their five dummy variables, plus one of the predictor variables of record. The restricted model was obtained by dropping the single variable of record data. Table 19 has been abridged by deleting predictors which showed no significant contribution for any AFSC. This resulted in dropping the scores from seven experimental tests: decoding, pursuit, figure analogies, cubes, mechanical principles, following directions, and practical estimations. The variable relating to size of the city of origin was dropped. Two Aptitude Indexes, the Administrative and Electronics, both of which had operated to select a sample, were not significantly predictive, and these were deleted. It should be borne in mind that the samples contained 399 observations per correlation coefficient, or more, and that the full model contained only 11 predictors. Consequently, the results are not particularly impressive, because small differences in prediction can be statistically significant. The spatial reasoning test score appears to have had some predictiveness for AFSC 291X0. Also, it is possible that the women switchboard operators contributed overall performance that were rated higher than those of other types of personnel. The complex scale reading test seems to have been predictive for personnel in the 304X4 specialty. Also, their status as divorced or separated was significant. However, this latter type of item depends upon very few cases, and the relationship could easily arise from chance associations due to selecting cases. Grade, experience, age, and maturity (all closely related factors) appear to have been predictive of overall performance in AFSC 431X1C. This is the type of relationship that one would expect in a specialty where most of the incumbents did a majority of the tasks. There was also a marginal indication that the memory for landmarks test was predictive. The two experimental test predictions for AFSCs 291X0 and 304X4 are provocative, but not sufficiently firm to be a basis for recommending a research program.

IX. SUMMARY OF REGRESSION PROBLEMS

The central issue of the regression problems was to determine if overall performance ratings could be satisfactorily predicted from data of record; that is, whether overall ratings could be predicted from such kinds of data as grade, Aptitude Indexes, age, education, and experimental test scores. The question was whether the individual task performance ratings predicted more of the overall variance than could be achieved by combinations of readily obtainable scores. An additional question concerned the specificity of the overall performance ratings for an AFSC: Was the composition of performance in the three AFSCs discernibly different? The within-rater agreement on overall ratings set an upper limit upon their predictability. Roughly, this was found to be r = .80, which translates into a 64 percent variance limitation on prediction. AFHRL obtained (in parallel analyses) cross-rater predictions that would translate into lower limits of prediction at 20 to 25 percent of the criterion variance.

In order to utilize the samples it was necessary to combine data from incumbents who were not rated on the same tasks. Also, in order to answer the questions about the relative contribution of the different kinds of data it was required that there be data of record on all ratees. The problem was approached by developing a "flagged" population whose overall criterion scores could be used as a standard of predictability. The samples taken for the regression equation could then be subsamples of the flagged population, and their representativeness could be tested.

The requirement for experimental test data narrowed the samples by at least half the cases. At the same time the need to assemble a true intercorrelation matrix forced the use of membership variables, and this doubled the number of predictors for task rating data. This simultaneous doubling of the number of task rating predictors and halving of the number of incumbents brought the ratio of predictors to

observations per correlation coefficient to a barely acceptable minimum. Although the prediction of overall performance variance was at least 25 percent, the statistical significance was marginal because of the large number of predictors.

The number of predictors was reduced to two sets of five task performance ratings, together with their dummy membership variables. Although some ratees may not have been adequately represented by task performance scores, the method provided an unqualified determination that as few as five task performance ratings could predict 20 percent of the overall performance variance. This prediction was compared to a maximum of 7 percent overall performance prediction from demographic or aptitude data, and to 5 percent unique prediction. That is, the data of record contained only a 5 percent determination of overall performance accountability which was unavailable from task performance ratings. In assessing the probability that task ratings were making a unique contribution to predicting overall performance ratings the findings were virtual certainty; the likelihood that the differences would have arisen by chance was less than one trial in 1,000 for every dimension of all three AFSCs.

Although the contribution of data of record was small, a few of the variables made significant increases in the total accountability of overall performance ratings. As one might expect, grade was the largest single contributor, accompanied by variables with which grade is correlated. Nevertheless, certain specific experimental tests were predictive, and the amount and significance of their predictiveness differed among the specialties.

X. DISCUSSION AND CONCLUSIONS

The highest correlations on task performance were found among the tasks of the 304X4 AFSC. It has been found possible to bring some information collected outside this study to bear on the observation. The comparative task difficulties for the three AFSCs are shown in Figure 7. Based on a scale of 7, and reflecting the mean of NCOs' ratings of the length of time it would take to learn each task as compared to others in the inventory, the distributions in Figure 7 show that the tasks of AFSC 304X4 were rated as much more difficult than most of those in the other AFSCs. About one-third of the tasks used in the performance study had to be omitted from this graph because they had been combined in a way that was not the same as the inventory. However, it was ascertained (by examination of the separate difficulty ratings of the tasks that the contractor had combined) that had the omitted tasks been included as given in the inventory the resulting distribution would have shown at least as great a difference from the others. When rating task difficulty, the NCOs were restricted to the inventory presented to them, and the data of the three graphs presented in Figure 7 have no common standard among AFSCs. Nevertheless there is reason to believe that a real connection exists between task difficulty and the rateability of task performance. This accords with the presumption that in order to rate task performance reliably there must be observable differences in task performances. It follows that the greatest differences in task performance should occur in those tasks which take the longest to learn. Two factors govern the observed data. One is that a gradation in performance must be present if one is to see that a task is not done perfectly; the other factor is a sampling consideration. At any one time of sampling performance there should be individuals at varying stages of learning. Neither gradations in performance nor occurrence of individuals at varying stages of learning is to be found for simple tasks of the done/not done variety, particularly if those tasks are sufficiently simple to be done on a single verbal instruction.

About 20 percent of all tasks shown in Figure 7 lie below the value 4 on the 7-point scale of difficulty. These are probably the type of task that is either done or not done, with little or no possible gradient in performance. Efforts are being made to investigate the nature of supervisor agreement on the performance of jobs with a preponderance of such tasks, since adaptability to jobs of that kind is related to job satisfaction and may throw light on the adaptability of certain kinds of personnel to military service. All of the foregoing emphasizes the desirability of discovering predictors of performance that are effective at the task level. Going beyond this, the observation that AFSCs with different difficulty levels are unequal

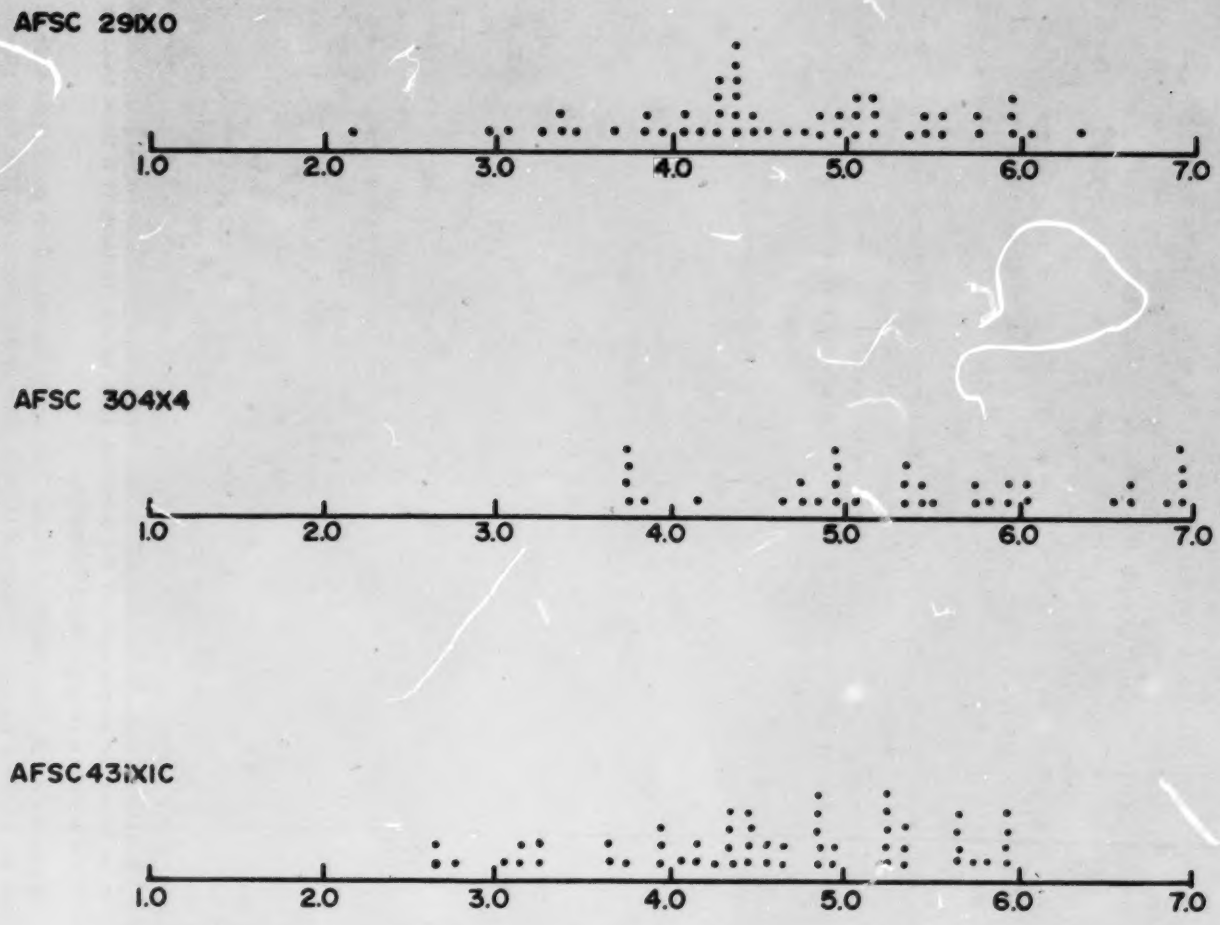

Figure 7. **Comparative Task Difficulty Means by AFSC.**

in performance prediction raises the question of differences of job types within an AFSC. Are job types unequal in performance evaluation when unequal in difficulty? Analyses are being made which will determine if the data collected in the present study are sufficient to throw light on this additional problem.

One must look with care at all predictors in the present study because the three AFSCs have extremely different aptitude cutoff points in Air Force training assignment. Owing to the high correlation among Aptitude Indexes the three samples are stratified, and this is apparent on all experimental test score means and on educational background. Despite the fact that the members of the 304X4 AFSC were elite with respect to aptitude, their overall performance rating means and their task performance rating means were lower than the other two AFSCs. Thus, an attempt to combine the three samples would have resulted in mixing incompatible criteria of performance. Not only were the performance criteria different for the three, the distributions of their predictor scores were unlike. Consequently, the AFSC samples must be treated as populations, and generalizations made beyond the framework of a single AFSC must be very cautious. While this somewhat limits the kinds of conclusions one might like to draw from the study, it is in itself an observation of value because the data reflect conditions that will reappear in the operational Air Force.

Small, statistically significant contributions toward predicting overall performance were made by specific experimental tests for some specialties but not others. The test predictions may be worth further exploration, but on a research basis to discover general areas of behavior that are not tapped by existing batteries. The practical effects on selection and assignment would be slight. Lest one become overenthusiastic, it should be remembered that a 2 percent increase in overall accountability would be statistically significant.

When one attempts to account for differences among the regression equation patterns predicting the various kinds of performance, he is faced with the distinctions in working conditions afforded by the three specialties. For example, the Aircraft Maintenance Specialist tended to have jobs which included a large number of the listed tasks in the performance survey. This was quite different from the jobs in the other two surveyed groups. Possibly the fact that a single Aircraft Maintenance Specialist job sampled a large number of the tasks caused a lack of specificity among the predictors, which resulted in an increase in the effect of grade and length of service on predicting overall performance ratings. Very discrete jobs marked the Telecommunications Operations Specialist assignment, and a wide variety of jobs could be found for Ground Radio Communications Equipment Repairmen. A continuous flow of work passes through a communications center, while the work of a repairman is marked by completion of labor on a tangible piece of hardware. The predictors having weight for overall dimensions may offer a clue as to how the work situation affects the human requirements, but at this point the conclusions would be largely speculation. For example, the dimension of quantity of work was better accounted for in the 304X4 AFSC than one might expect; but, on reflection, it seems possible that quantity of work in this case is a function of the rate at which a piece of equipment is adjusted, which is a matter of skill rather than motivation and energy. The results of this portion of the study offer attractive leads, but they are not definitive.

The enterprise has reaffirmed the thesis that the assessment of job performance is a complex process. It has suggested that the factors involved may not be the same for different AFSCs. One of the strongest factors which distinguish jobs is the relative difficulty of their tasks. More difficult tasks were associated with more measurable performance differences. This was evidenced by better agreement between pairs of raters, less piling up of ratings at the top of the scale, and better agreement between task performance ratings and ratings of whole job performance. Nevertheless, the performance of incumbents in jobs made up of less discriminating tasks was still differentiated. In such cases it seems likely that other aspects of behavior operated to distinguish performance, possibly factors of motivation and attitude. Demonstration of that hypothesis must await future analyses of these data.

From its outset, the study was known to be a high-risk enterprise; a serious likelihood existed that no definitive findings would be made. Fortunately, clearcut observations have been established. Three of these appear to be crucial: (a) performance data can be collected at the level of the task, which is the basic unit

of a job inventory; (b) differences in task performance are observable for those tasks that NCOs rate as being difficult in terms of required learning time; and (c) excepting behaviors exhibiting poor motivation, persons who are faced with rating performance of very simple tasks tend to credit the worker with perfect performance. This doesn't seem to be a great deal to extract from such a large enterprise. Could not almost anyone have reasoned out the last two conclusions without the study?

The answer is yes, maybe. The point is that the literature on job performance is packed with observations about inflation of ratings, but very little has been offered which associates inflation with an honest inability to discriminate between performances. The implications are not trivial. Coupled with the finding that there are tasks whose performances can be discriminated, it is a finding of immediate value to occupational analysis. It removes a very considerable volume of tasks from consideration in terms of aptitude and lengthy training, and it strengthens the need to concentrate research and development efforts on tasks that have been selected for their criticality.

What is this "criticality"? It seems to be a two-element concept. The first is that jobs are critical if failure to perform them properly has disastrous results. The second element is the relative difficulty of filling the jobs. Since all jobs contain many simple tasks as well as complex and difficult ones, the composition of a critical job is important. First, it is necessary to determine which tasks make the job critical in the sense that the tasks must be properly done. Next, it is necessary to determine if the critical tasks are actually difficult. Finally, if the objective is to select personnel to fill those jobs, it is necessary to find an adequate sample of already filled jobs to use as performance validators.

The last requirement is often very difficult to meet. Jobs containing the critical task may be, and usually are, in small numbers. Incumbents will report varying proportions of their time spent on the tasks which are of interest for training and assigning personnel. Consequently, one must be assured in picking a sample of performance validating jobs that he is avoiding cases that were overrated because the incumbents spent very little time on the critical tasks.

The findings of the study point to several approaches which can be used in solving the validator problem. The most attractive of these capitalize on the fact that task difficulty meansures and performance discrimination data can be combined with job clustering techniques. One such method might be to combine job inventory responses with Automatic Interaction Detector (AID) (Koplyay, Gott, & Elton, 1973) to determine the relevance of certain tasks, or task clusters, to rated overall performance. The point is that we have the statistical methods at hand to effect improvements in identifying future assignees to critical jobs. The results of this study are encouraging in that they indicate the feasibility of obtaining the requisite data.

There is a grey area in the analyses of these data which may turn out to be important for practical purposes. Statistically significant differences were found in performance assessments of incumbents with relatively simple tasks. We have been able to bring to bear only one dimension of task characteristics as a work requirement, NCOs' task difficulty rating. As a result of this limitation we have assumed that some of the unmeasured performance variance was attributable to attitude and interest. It has not been possible in the initial analyses to break this category down, and as a result the explanation may be an oversimplification. One can cite the requirement for supervisors to instruct subordinates, and the need to exercise interpersonal skills in the process. One can also cite the possibility of lowered performance under conditions of stress and emotional tension, which might affect the execution of very simple tasks. This is related to work that is currently going forward in occupational analysis on the development of benchmark scales for several dimensions of task and job performance (Goody, 1976).

XI. SUMMARY

A study was organized to evaluate the potential usefulness of rating an airman's performance on individual tasks. Broadly, the purpose was to measure job performance, but more particularly, to assess the contribution of task performance ratings to such measurement. Making assessments at the task level has the practical advantage of using that unit of measurement on which job inventories are constructed and analyzed.

The study was necessarily an ambitious one because of the multitude of variables that might contribute to performance ratings, whether they are task performance ratings or overall evaluations. Moreover, the study had to be representative of Air Force specialties and the tasks within those specialties. While collecting data in depth it was reasonable to collect corollary information which might contribute to understanding the findings. Thus, data were collected on sources of training on tasks, on the motivational aspects of doing tasks, and on retention of skills. The underlying aspects of evaluation were sought, both at the task and at the global levels, by obtaining self-ratings on task performance from incumbents, which were paralleled by incumbent assessments of the tasks as motivators, and from peers and supervisors, who rated task performance, task performance moderated by motivation and training, and six kinds of overall performance. The incumbent data bank included a current job inventory, and a test battery containing both experimental cognitive tests and measures of attitude, job satisfaction, and interests. Air Force files were matched to retrieve such information as grade, education at entrance to service, age at entrance to service, total service time, and Aptitude Index scores. After considerable development effort utilizing NCOs as consultants, the following specialties were chosen: AFSC 29150, Telecommunications Operations Specialist; AFSC 30454, Ground Radio Communications Equipment Repairman; and AFSC 43151C, Aircraft Maintenance Specialist, single- and dual-engine jet.

Multiple goals were listed in the proposal of the study, and in recognition of the immensity of the data reservoir, the accumulated information was also regarded as a data bank for exploration of promising leads, and for aid in the solution of continuing methodological problems. A partial list of the goals includes:

1. Determining if raters can agree on an airman's quality of performance on individual tasks;
2. Searching for aptitude measures not in current use which are predictive of task or job performance ratings;
3. Evaluating existing aptitude measures as correlates of performance;
4. Determining which items of record, such as grade, account for assessed performance;
5. Assessing motivation and measuring its contribution to performance evaluations;
6. Assembling numerous measures to weight their relative importance in accounting for work performance.

Circumstances at the time of the study threatened to reduce the number of cases collected below the level essential to studying the performance of individual tasks. A supplemental survey was made with only one supervisor as a rater in addition to the incumbent. This prevented loss of the critical aspect of the study, task performance assessment, but it precluded some of the peripheral analyses. Among the casualties were job inventory data, which contained information on time on job, and a set of work requirement ratings applied to whole jobs. Air Force enlisted discharges proceeding during that period were probably the cause of reduced returns on a Time 2 evaluation obtained from supervisors. Despite all of this, the major purposes of the study were met, and the analyses resulted in very positive findings.

The basic findings can be listed as follows:

1. Raters can agree on task performance evaluations to a statistically significant degree.
2. Raters agree better on rating overall job performance than on rating the performance of single tasks.
3. Incumbents are poorer sources of task ratings than peers or supervisors.
4. Peers can be substituted for supervisors as performance evaluators without great loss in reliability.
5. A few task ratings taken together account for a substantial percentage of the overall performance rating variance.
6. Either, or both, aptitude data and demographic data (such as grade and length of service) account for much less of the overall evaluation of an airman's performance than can be had from ratings on as few as five tasks.
7. It is likely that some of the unaccounted-for overall performance variance was attributable to the attitudes and interests of the incumbents. (Only preliminary analyses could be made of these data, but the prediction results were statistically significant.)

8. It is possible that small improvements can be made in selecting personnel for certain task assignments by means of tests not in current use; but these improvements will be slight, at best.

9. Marked differences distinguished the 304X4 AFSC performance ratings from the other two specialties; they were internally more reliable and better able to predict overall performance ratings.

10. Use of the top of the rating scale was frequent when performance ratings were made on easier tasks, and on performance in AFSCs with lower aptitude requirements.

11. Since measurability was better for more difficult tasks, with less use of the top of the rating scale, the AFSC with high aptitude incumbents received the lowest mean performance ratings of the three specialties.

The importance of the last finding is only beginning to be appreciated. It suggests that very careful examination should be made of the task composition of jobs in those specialties that have a substantial range in the difficulty of tasks. In a sense, performance measurability becomes synonymous with task difficulty as it is being operationally recorded for Air Force job inventorying. This provides a preliminary index of tasks that can be used to identify difficult jobs. It also provides a warning that job performance data taken without regard to the task difficulty composition of jobs used as validators of selection procedures can yield misleading data. It would seem that the generality of this finding ought to apply outside the Air Force as well as in it. If this is true, it is likely to produce reassessments in many areas of validation for selection, classification, and job satisfaction prediction.

REFERENCES

Goody, K. *Task factor benchmark scales for training priority analysis: Overview and developmental phase for administrative/general aptitude area.* AFHRL-TR-76-15, AD-A025 847. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976.

Hahn, C.P. *Development of task level job performance criteria interim report, Phase I.* Washington, D.C.: American Institutes for Research, February 1973.

Hahn, C.P. *Development of task level job performance criteria final management report,* AD-A051 958. (Appendix A, AD-A051 959; Appendix B, AD-051 960; Appendix C, AD-A052 157; Appendix D, AD-A051 961.) Washington, D.C.: American Institutes for Research, 1975.

Koplyay, J.B., Gott, C.E., & Elton, J.H. *Automatic interaction detector-version 4 (AID)-4 reference manual.* AFHRL-TR-73-17, AD-773 803. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1973.

Wiley, L.N. *Airman job performance estimated from task performance ratings.* AFHRL-TR-76-64, AD-A034 320. Lackland AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, October 1976.

NOTES AND STUDY NUMBERS

Work was performed by AFHRL's Computational Sciences Division in the production of the Contract Monitor's contribution to this report. Due to the special circumstances under which work has been continued with these data and the ongoing nature of the work, it is thought desirable to distinguish the previous analyses from any that may have been commenced concurrently, but which have not been released.

Computational Sciences Division, SM, Study Number	Title and Comments
6284	Cross-Rater Task Prediction of Overall Ratings, TSKLVL, consisting in the samples reduced to incumbents with both raters, mostly peer and supervisor. The two sets of 5 task performance ratings by one category of rater were used in a test to predict the overall ratings made by the other category of rater.
5930	Pay-off Regressions for AIR TSKLVL Final Report. Regression problems predicting overall criteria, with F-tests by blocks of predictors.
5734	Task Performance Predictions of TSKLVL Overall Criteria. Intercorrelations with membership variables for task performance ratings with overall criteria using "flagged" samples.
5687, 5668, 5530	Title not relevant. These were preliminary analyses involving regression prediction of overall performance from cognitive tests and arbitrarily weighted noncognitive tests. These formed the basis for concluding that attitude data would contribute to predicting overall performance, differentially by AFSCs. Flagged samples had not been developed, and further work would use flagged samples when completed.
5196	Consolidate F41609-71-C-0010 Survey Data. These are the basic record tapes containing the major protocols and the test data for the entire study. Matching data from Air Force records are contained under other study numbers.
5135	Shakedown Analysis of F41609-71-C-0010 Tapes. Correlations and intercorrelations used to check contractor's accuracy and quality, along with such preliminary analyses as seemed useful.

☆U.S. GOVERNMENT PRINTING OFFICE:1978- 771-122/24